Data Remanence in New Zealand

D. Roberts

A thesis submitted for the degree of

Doctor of Philosophy

at the University of Otago, Dunedin,

New Zealand

7 March 2013

Abstract

International research has shown that individuals and companies in other countries do not always fully remove the data from their computer data storage devices before disposing of them. Typically this means when people are disposing of their computer hard drives there is a wealth of personal or corporate information that can be exploited to commit such crimes as identity theft, fraud, stalking and blackmail.

A further literature review showed that no such "data remanence" research for hard drives (or any other data storage devices such as mobile phones, USB thumb drives and the like) had been conducted in New Zealand.

The methodologies for all relevant hard drive data remanence experiments were compared and then used to design the most appropriate methodology for this research.

100 second hand hard drives were then sourced nationally across New Zealand for the experiments of this research to determine the baseline of data remanence for hard drives in New Zealand. The results of the experiments were then compared with international results to determine how New Zealand compares and what if any further actions (such as education) should be taken.

Acknowledgements

Writing the acknowledgements is often one of the hardest parts of a thesis. There are simply too many people to thank and no really good scheme to organise those acknowledgements. If I have missed you out, please accept my apologies.

I need to thank:

My parents and grandparents who were great role models and fostered a love of learning that eventually led to this thesis.

My primary supervisor Dr Wolfe for his years of support, guidance, wisdom and sense of humour.

The Infosci office staff, especially Gail Mercer and Stephen Hall-Jones.

The Infosci Tech Support Group for always having the right equipment or the right solution to the thorny hardware problems encountered in this thesis.

My proofreaders Dr Melanie Middlemiss and Miss Amberleigh Nicholson for their excellent eye for detail, constructive criticism and spotting the really obvious mistakes tht always creep in.

Special thanks go to Dr Nigel Stanger for his feedback and thought provoking suggestions that really improved this thesis and the future work that will come from it.

Special thanks also to Brian Niven of the Maths and Stats department for his excellent advice and recommendations regarding statistics in this thesis.

The University of Otago for funding parts of this research.

The thesis examiners and Dr Martin Purvis as the convenor for providing timely and important feedback that improved this thesis.

Dr Chris Roberts* and his co-workers for recognising the importance of this topic and encouraging me to research it.

There were so many friends, office mates and fellow students who offered a laugh, distractions, shenanigans or really good ideas for this thesis.

Some of those people (in alphabetical order) include: G. Can, Gerald Cochlan, G. Corso, Kris Evans, Jason Glazier, A. Godek, K. Gorton, “K.H”, J. Horton, Mark Hodge and members of the various board gaming groups, T. Inley “S.D.K”, L. Lau, Simon Mooney, Brendan Murray, J. Norris, Kevan Quinn, Bobbi Smith, Clint Sparks, and Ed Stackhouse.

**No relation*.

Table of Contents

Abstract ii
Acknowledgements iii
Table of Contents iv
List of Figures ix
List of Tables x
Chapter 1: Introduction 1
1.1 Introduction 1
1.2 Key Definitions 2
1.3 Research Objectives and Methodology 2
1.4 Contributions 3
1.5 Structure 4
Chapter 2: Background 5
2.1 Introduction 5
2.2 Data Storage Devices 5
2.2.1 Physical hardware of Hard Drives 6
2.2.1.1 Electro-mechanical Drives 6
2.2.1.2 Solid-State Drives 8
2.2.1.3 Hybrid Drives 8
2.2.2 Interface Types 9
2.2.2.1 Small Computer System Interface (SCSI) 10
2.2.2.2 Integrated Device Electronics (IDE) aka ATA 11
2.2.2.3 Comparison between SCSI and IDE 12
2.2.2.4 Serial AT Attachment (SATA) 13
2.2.2.5 Universal Serial Bus (USB) 14
2.2.2.6 Adapters 15
2.2.3 Interactions between Hard Drives and Operating systems 15
2.2.3.1 Hidden Disk Areas 15
2.2.3.2 Master Boot Record 16
2.2.3.3 Partitions 16
2.2.4 Other data storage devices: Mobile/Smart Phones 17

2.3 Technical issues regarding Hard Drive Data Remanence....21
2.4 Forensic Computing....24
2.4.1 What is Forensic Computing....24
2.4.2 Minimal handling of the Original....24
2.4.2.1 Dead forensics....25
2.4.2.2 Live forensics....26
2.4.3 Accounting for Change....26
2.4.4 Comply with the rules of evidence....28
2.4.5 Implications for this research....29
2.5 Psychology of Security....29
2.5.1 Relevance to Data Remanence....32
2.6 Risks arising from Data Remanence....32
2.6.1 Legal Requirements: New Zealand Privacy Act....33
2.6.2 Espionage....34
2.6.3 Identify Theft....35
2.6.3.1 Financial Identity Theft....35
2.6.3.2 Non Financial Identity Theft....37
2.6.3.3 Criminal Record Identity Theft....37
2.6.3.4 Potential Identity Theft and Data Remanence....37
2.6.4 Blackmail....39
2.6.5 Stalking....39
2.7 Summary....39
Chapter 3: Literature Review....40
3.1 Introduction....40
3.2 Pilot studies....40
3.3 Longer term and comparable research....43
3.4 Specialised research questions....45
3.4.1 Primary Health Information....45
3.4.2 Identity Theft....47
3.4.3 A mixed approach....48
3.5 Data remanence for other data storage devices....49
3.5.1 Mobile/Smart phones....49
3.5.2 USB storage devices....50

3.6 Techniques to automate Data Remanence research 51
3.7 Auxiliary research 53
3.8 Literature review summary 53
Chapter 4: Methodology 55
4.1 Introduction 55
4.2 Special considerations 55
4.2.1 Ethical Approval 55
4.2.2 Special protocols for objectionable material 56
4.2.3 Securing the data 57
4.3 Methodology design 58
4.3.1 Comparison of methodologies from prior researchers 58
4.3.2 Other considerations from previous methodologies 60
4.3.2.1 Reporting 60
4.3.2.2 Sample size considerations 60
4.3.2.3 Research questions 62
4.3.2.4 Identifying Information 62
4.3.3 Experimental procedure 63
4.4 Documentation 66
4.5 Sources of Hard Drives 67
4.5.1 Auction sites 68
4.5.2 Private companies that sell second hand hard drives 70
4.6 Acquiring the data 71
4.6.1 Hardware used 71
4.6.2 Determining whether a drive was unreadable 73
4.7 States of data on readable drives 75
4.7.1 Wiped 75
4.7.2 Reformatted 76
4.7.3 Clearly viewable 76
4.7.4 Special case: Encrypted 77
4.8 Automating the research using the Spike Effect 77
4.8.1 Results of running the Spike Effect program over specific drives 80
4.8.2 A heuristic to determine whether a drive has been wiped or not. 85
4.8.3 Comparing the Spike Effect program with *FTK* 85

4.8.4 Developing a Spike Effect program variant that uses sampling 87
4.8.4.1 Systematic Sampling 87
4.8.4.2 Simple Random sampling 91
4.8.4.3 Other potential approaches instead of sampling 92
4.8.5 Future Work for the Spike Effect 92
4.9 Analysing non-wiped drives 92
4.9.1 Choice of analysis tool 92
4.9.2 Determining which data to analyse 94
4.10 Reporting the data found 96
4.11 Methodology Summary 99
Chapter 5: Analysis of the Data 100
5.1 Introduction 100
5.2 Sources of the drives 100
5.3 Observations and changes from the methodology 101
5.4 Brief Results 101
5.5 Hard drives of interest 102
5.6. Analysis of results by interface type 104
5.6.1 Introduction 104
5.6.2 Unreadable drives by Interface: 105
5.6.3 Wiped Drive by Interface 106
5.6.4 Interface comparison by Identifying information 107
5.6.5 Summary of comparisons by interface 108
5.7 Analysis by source 109
5.7.1 Introduction 109
5.7.2 Unreadable drives by source 109
5.7.2 Wiped drives by source 111
5.7.3 Identifying information by source 112
5.7.4 Analysis of Company A versus Company B 112
5.7.5 Summary of impact of the drive source 113
5.8 The Primary Research Objective 113
5.8.1 Comparison of New Zealand results and Consortium combined results 114
5.8.2 New Zealand results compared versus individual consortium members 116
5.8.3 Summary of the primary objective 116

5.9 Secondary Research Objectives 116
5.9.1 Company Types 117
5.9.2 *Tableau* investigation 120
Chapter 6: Conclusions 123
6.1 Introduction 123
6.2 Primary Research Objective 123
6.3 Secondary Research Objectives 124
6.3.1 Company Types 124
6.3.2 "How reliable is the *Tableau*'s 'Source may be blank' function?" 125
6.4 Future work 127
References 129
Glossary 136
Appendices 139
Appendix A Price Comparison of Hard Drives. 139
Appendix B Output of Spike Effect on Drive SCSI_03 141
Appendix C Full results of the experiments 144
Results from Company A : Drives 1 to 50 144
Results from Company B : Drives 51 to 69 149
Results from Trade Me : Drives 70 to 100 152
Appendix D Ethical approval for this research 158
Appendix E Valli's initial methodology 159
Appendix F NIST definitions 160
Appendix G Vendor description of *Tableau*. 162
Appendix H ANZSIC LEVEL 1 CLASSIFICATIONS 163
Appendix I Consortium results 2006-2009 164
Appendix J Spike Effect Systematic Sampling processing times 166
Appendix K Ranked comparisons of data remanence 170

List of Figures

Figure 1 Electro-mechanical Hard Drive example (Surachit, 2007) 6
Figure 2 Flowchart for classifying hard drives 65
Figure 3 Example of "Totally wiped" drive 81
Figure 4 Example of "Practically wiped" drive 81
Figure 5 Example of "Very diverse data" drive 82
Figure 6 Example of a "Special Case" drive 83
Figure 7 Master Boot Record from SCSI_03 83
Figure 8 Comparison of the four drive types 84
Figure 9 *Forensics Tool Kit* (*FTK*) screenshot 93
Figure 10 Data carved Flag 94
Figure 11 Flag of New Zealand 95
Figure 12 Flag of Australia 96
Figure 13 Comparison of processing times for IDE_39 166
Figure 14 Comparison of processing times for IDE_44 166
Figure 15 Comparison of processing times for IDE_46 167
Figure 16 Comparison of processing times for SCSI_03 167
Figure 17 Comparison of processing times for IDE_71 168
Figure 18 Comparison of processing times for IDE_72 168
Figure 19 Comparison of processing times when the batch script has been changed 169

List of Tables

Table 1 Amalgamation of consortium 2006-2009 results....44
Table 2 Comparison of different features from published methodologies....59
Table 3 Sample Size Example Table....61
Table 4 Comparison of Unreadable Drives from Consortium Results 2006-2009....73
Table 5 Example of the Spike Effect program processing time for IDE_44....85
Table 6 Comparison of Spike Effect program and *FTK*....86
Table 7 Counts of Spikes found when sampling IDE_39....89
Table 8 Counts of Spikes found when sampling IDE_72....90
Table 9 Counts of Spikes found when sampling SCSI_03....90
Table 10 Example of Consortium reporting....97
Table 11 Template table for answering "How many drives were wiped?"....98
Table 12 Template table for answering "How many drives had identifying information?"....99
Table 13 Total of all New Zealand Hard Drives analysed....101
Table 14 Breakdown of Identifying Information as found on New Zealand Hard drives....102
Table 15 Summary of Interface by Source....105
Table 16 Summary of Unreadable drives by Interface....106
Table 17 Comparison of Interface type on wiped drives....107
Table 18 Comparison of Interface and Identifying Information....107
Table 19 Summary of drive source and drive state....109
Table 20 Summary of drive source and identifying information found....112
Table 21 Direct comparison New Zealand vs. Consortium 2009....114
Table 22 Cross tabulation of wiped drives for New Zealand and the consortium....114
Table 23 Cross tabulation for company identifying information comparison....115
Table 24 Breakdown of Company Identifying Data found.....118
Table 25 Table by Type of Industry....119
Table 26 Spike Effect output for "Source Drive may be Blank"....121
Table 27 Comparison of Hard Drive prices....140
Table 28 2006 Consortium Results (Jones, et al., 2006)....164
Table 29 2007 Consortium Results (Jones, Valli, Dardick, et al., 2009)....164
Table 30 2008 Consortium Results (Jones, Dardick, et al., 2009)....165
Table 31 2009 Consortium Results (Jones, et al., 2010)....165
Table 32 Rankings based on Unreadable Hard drives....170
Table 33 Rankings based on Wiped Hard Drives....170
Table 34 Rankings based on Company Identifying Information....171
Table 35 Rankings based on Individual Identifying Information....171

Chapter 1: Introduction

1.1 Introduction

Recent newspaper headlines include:

"Missile data, medical records found on discarded hard disks"

(Leyden, 2009)

"Dumped computers exploited by crims"

(Stuff.co.nz, 2011a)

"Old hard drives a fraud risk"

(Stuff.co.nz, 2011b)

Reading the stories behind the headlines it is clear that discarded or second hand computer data storage devices (hard drives, USB drives, and cell phones, for example) can present a variety of potential risks to their former owners (individuals or companies) and also for any people whose information was stored by those former owners. New Zealanders may face the same risks that Australians and Britons face from disposing of their computer hardware, but a literature review shows that while there has been ongoing research in this field (of "data remanence") since 2005 for Australia, the United Kingdom (UK) and other countries, similar research has not been published about New Zealand.

As there is no baseline for the level of data remanence in New Zealand this thesis will examine New Zealand data storage devices (in this case computer hard drives) to determine firstly what the level of data remanence is and secondly if the results of research from other countries is directly comparable or not.

Data remanence can be defined as "Data remanence is the residual physical representation of data that has been in some way erased. After storage media is erased there may be some physical characteristics that allow data to be reconstructed." (National Computer Security Center, 1991) Defining "the level of data remanence" can be difficult. This is because different researchers use different metrics and reporting styles for their results. For example, the level of data remanence could be measured as the number of devices out of the sample

size that have had all the data permanently removed (wiped) from them. Another metric could be counting the specific number of files found such as "Hard drive X had 2918 image files and 3100 word document files". This thesis will examine a number of reporting styles, metrics and methodologies used in data remanence research and determine an appropriate methodology for measuring and reporting the baseline for New Zealand.

1.2 Key Definitions

There are a number of terms that will often be used in this research.

Company: This typically refers to a business, but it may also include not-for-profit organisations, educational institutions such as high-schools, and also government agencies.

Consortium: The term "consortium" is used to refer to the group now known as the "Cyber Security Research Network" that involves a group of data remanence researchers from Australia, France, Germany, the United Kingdom and the United States of America.

Data remanence: Data that physically remains on a data storage device after it has been prepared for disposal.

Typical/Home User: A person with below average to average computer skills using their computer in a home setting, rather than at a work place where someone else is likely responsible for the computer's ongoing care and disposal.

A full glossary can be found at the end of this thesis (starting at page 137).

1.3 Research Objectives and Methodology

The primary objective of this research is to determine the baseline of data remanence in New Zealand. This will assist in determining trends in data remanence in New Zealand in future work and will also allow for comparisons between New Zealand and other countries such as the consortium in data remanence. The secondary objective is to determine if certain types of company in New Zealand are at greater risk from data remanence than other companies.

A second secondary research objective is to develop an analysis technique to determine if hard drives have been wiped which may assist in automating forensic analysis. This objective was also used to determine "How reliable is the *Tableau*'s 'Source may be blank?' function?"

The methodologies from previous research such as the consortium and others will be compared to find the most practical methodology for studying data remanence in New Zealand. Examples of data that researchers have previously looked for include medical records, data that facilitate corporate espionage, or data that facilitate identity theft (El Emam et al., 2007; Jones, 2005; Medlin and Cazier, 2011).

Often but not always researchers have examined the remnant data for information both about individuals and companies. This research will therefore also examine hard drives sourced from across New Zealand for any company identifying information and attempt to determine which types of New Zealand companies, if any, are at more risk.

1.4 Contributions

The primary contributions from this research were presented at the 9th Australian Digital Forensics Conference (Roberts and Wolfe, 2011).

The contributions of this research are:

1. Determining the baseline of the level of data remanence in New Zealand for hard drives sold on the second hand markets.
2. Comparing the baseline with other countries to determine if New Zealand was any better or worse with regards to preparing hard drives for disposal.
3. Determining whether certain company types in New Zealand are greater risk from data remanence than others, which other researchers have not generally considered for their own countries.
4. Developing an analysis technique to determine if hard drives have been wiped that can assist automated forensic analysis and save overall processing time.
5. Determining the reliability of the *Tableau TD-1* "Source drive may be blank" functionality.

The fourth and fifth contributions have not been published outside of this thesis as the tool to implement the analysis technique is undergoing further development. It is expected the tool will be released in either software or hardware form when it has undergone suitable testing (potentially to NIST standards).

1.5 Structure

The remainder of this thesis is structured as follows.

Chapter 2 presents a background of data remanence issues including the technical reasons for data remanence as well as human computer interaction issues, which can include the potential risks from data remanence.

Chapter 3 presents the literature review. It considers the pilot studies then the continuing research. The literature review then examines specific questions that various researchers have researched with regards to data remanence. The majority of the data remanence literature prior to 2011 dealt with hard drives sold on the second hand market.

Chapter 4 presents the methodology for this research. It initially considers the methodologies of other researchers as presented in Chapter 3, considers different ways of conducting and presenting the results of the research. Chapter 4 also introduces a new analysis technique "the Spike Effect" to assist in conducting data remanence research.

Chapter 5 presents the analysis of the experiments and considers those results within the New Zealand context and when compared internationally.

Chapter 6 presents the conclusion and discusses the results, their implications, and potential solutions and then discusses future work from this research.

Chapter 2: Background

2.1 Introduction

This chapter examines three main areas of background literature with regards to data remanence. Those areas are: the technical aspects of data storage devices that make them prone to data remanence issues, forensic computing, and the psychology of security.

2.2 Data Storage Devices

The main focus of this research is on the remaining data found on hard drives sold on the second hand market. Other data storage systems have been briefly included as the current technological issues surrounding performing data remanence research on those systems may be solved in the future.

This section introduces key concepts of data storage devices and then considers specific issues of data storage devices with respect to data remanence research.

Section 2.2.1 considers the physical hardware of a hard drive and describes electro-mechanical, solid-state and hybrid hard drives.

Section 2.2.2 considers the interface types of hard drives (IDE/ATA, SCSI and the like) and discusses why interface types were created, the history of hard drives before interfaces and their relevance to data remanence research.

Section 2.2.3 considers the interactions between the hard drive and the operating system. This incorporates concepts such as the Master Boot Record, partition tables and hidden disc areas such as the Host Protected Area and Device Configuration Overlay.

Section 2.2.4 considers the research and technical issues relating to other data storage devices such as mobile/smart phones.

2.2.1 Physical hardware of Hard Drives

This section gives a partial history of hard disk drives (HDDs) and the way they are manufactured. At the time of writing there are three main types of hard drive when classified by the methods used to store the data: Electro-mechanical, Solid-State and Hybrid.

2.2.1.1 Electro-mechanical Drives

The technology for electro-mechanical data storage has been available since 1899 when Poulsen invented magnetic recording by applying two principles (Piramanayagam, 2007, pg. 4):

1. Magnets produce strong magnetic fields at the poles. This field can be used for reading information.
2. The polarity of the magnet itself can be changed by applying external fields.

Electro-mechanical hard drives are mainly composed of two parts, the read/write head (the mechanical part of the name) and the storage medium or surface (often referred to as the platters) that stores information in binary format (1s and 0s). Figure 1 presents a modern electro-mechanical hard drive that uses the IDE interface (as shown by the IDE connector). The IDE interface as explained in Section 2.2.2.2 on page 11.

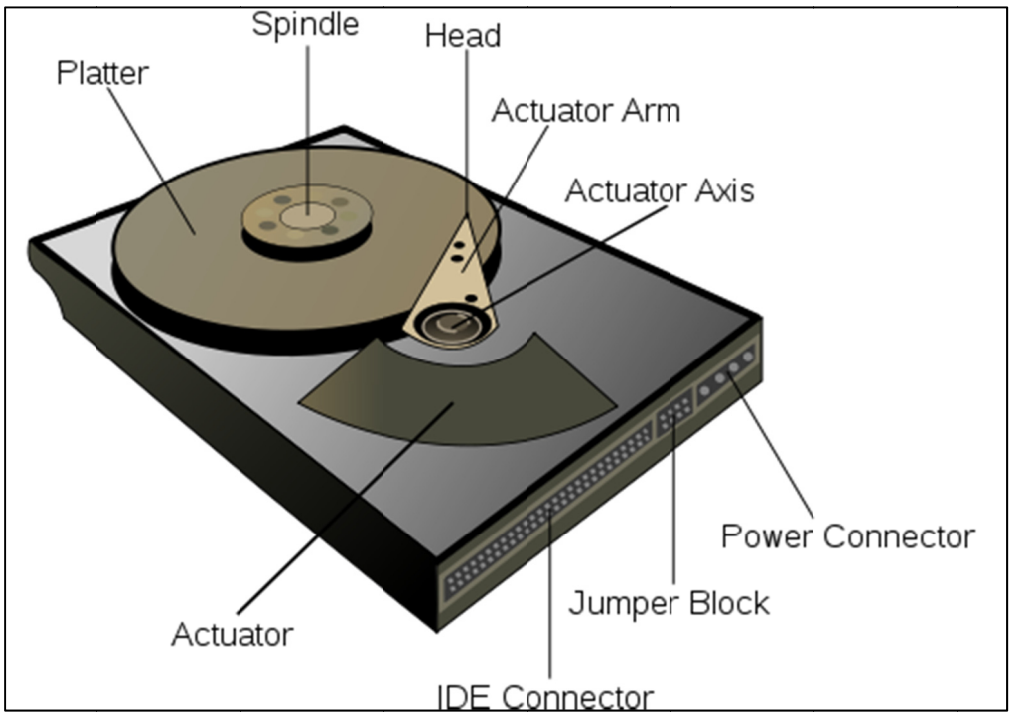


Figure 1 Electro-mechanical Hard Drive example (Surachit, 2007)

The important thing to note with Figure 1 is that while the platter, spindle and actuator components appear to sit on the top of the hard drive they actually reside inside the hard drive. The only external parts are the "IDE Connector", "Jumper Block" and "Power Connector". As per the description of how electro-mechanical hard drives work, the platter(s) rotate around the spindle and the actuator assembly either reads or writes the data as appropriate.

Anderson (2003, pg. 20) explains the history of hard drives by relating how magnetic tape devices were previously used which was linear (one-dimensional) and the change to a "three dimensional storage model" as involving a cylinder number, a head number, and a sector number. Anderson further describes it as being the original programming model for disk drives the "cylinder-head-sector access" (Anderson, 2003, pp. 20-21). The sector is further defined as the smallest addressable location on a hard drive and is typically 512 bytes in size.

The original hard drives from the late 1950s typically consisting 50 individual disks/platters of 24 inches in diameter and stored a total of 5 megabytes of data (Piramanayagam, 2007, pg. 2). In comparison, hard drives in 2007 were typically one or two platters of 3.5 inch diameter that stored approximately 750 gigabytes (Piramanayagam, 2007, pg. 2). This shows the advances in technology, that while the hard drive platter size has decreased by 85% (from 24 inches to 3.5 inches or smaller) the data capacity has increased by 150 000 times over 50 years (Piramanayagam, 2007, pg. 2).

Electro-mechanical drives can be susceptible to a number of faults. Those faults include "mechanical failure", "physical damage", "chemical", "electrical" and "firmware/ service area". Of note is one of the examples given under "mechanical failure" which is a "head crash". This can be caused by insufficient airflow to maintain an air bearing between the head and the disk platter or due to a sudden impact during operation or a power failure (Sutherland et al., 2010, pg. 158). Similar damage can occur when hard drives are transported without appropriate packaging which is likely to be a factor in data remanence research that relies on second hand markets.

2.2.1.2 Solid-State Drives

Chen et al. (2009) compare two forms of solid-state technologies with electro-mechanical drives. They describe solid-state disks (SSDs) as based on semiconductor chips which enables low power consumption and means the drives are compact, and more shock resistant than electro-mechanical hard drives (Chen et al., 2009, pg. 181). This suggests solid-state drives sold on the second hand market would be more likely to work compared to electro-mechanical drives.

One of the downsides to SSD versus HDD is the cost per gigabyte. Chen, et al. (2009, pg. 184) state that low end SSDs cost approximate US $5 per gigabyte while high end SSDs can cost as much as US $25 per gigabyte. In comparison, electro-mechanic hard drives cost only a few cents per gigabyte. As part of research for this thesis, sources and prices of SSD versus HDD technology were examined and the price disparity was still similar. Appendix A shows a sample of the different prices of storage technologies.

The design for solid-state drives deliberately incorporated backwards-compatibility with the existing interfaces used by electro-mechanic hard disk drives.

2.2.1.3 Hybrid Drives

As the name suggests, hybrid drives are a combination of both electro-mechanical and solid-state drives. Min and Nam (2006) describe hybrid drives as adding a flash memory chip to a (electro-mechanical) hard disk drive for reduced power consumption and faster start-up. In the example used in Appendix A, the hybrid drive listed had 500 gigabytes of electro-mechanical storage, and 4 gigabytes of solid-state storage. These drives attempt to use the best of both worlds of the fast processing times for the solid-state component and the very cheap bulk storage of the electro-mechanical component.

2.2.2 Interface Types

For the purposes of this section, "interface" refers to the connection between the hard drive and the physical computer. When discussing SCSI and other interfaces there may be references to the "bus architecture" which considers the entire design of a system rather than just hard drives. That "bus architecture" and similar discussion while interesting are outside the scope of this work.

To understand why storage devices have different interfaces, it is important to consider the history of hard drives. Anderson (2003) explains that prior to the SCSI and IDE interfaces being created, the operating system (OS) had to do all the work and understand each type of hard drive that could be connected to it. This means the host operating system had to know where the heads were for reading and writing, send the signals, perform checking and handle errors itself. This in turn required more resources, and an understanding of exactly how to do that for each different type and brand of hard drive inside the computer. "The OS or controller needed to include a table of drive geometries, the size of which was multiplied by the number of generations it contained." (Anderson, 2003, pg. 23).

With more manufacturers, different capacities and geometries the operating system needed to store even more information about all the potential different hard drives it could have attached to it. Conceptually this is similar to imagining that each manufacturer spoke a different "language" and the operating system would need to know each language and each regional dialect of that language.

In non-technical terms this is analogous to air traffic control systems. For example the situation prior to the creation of hard drive interfaces it would be like air traffic controllers having to know all the languages that any pilot entering their airspace knew. To make things safer and easier the decision was made that all pilots and all air traffic control staff must speak the common language of English on request. Functionally that is how interfaces came about, as a mechanism that meant the operating system could communicate in one "language" and the interface would translate it into the specific "language" that the hard drive used. The interface would take a standard "write data" command and would look after the details of positioning the heads to write data, rather than the host having to do this. This in turn would free up the host to do other things.

Benefits of having an "intelligent interface" such as SCSI and IDE/ATA include "sophisticated error recovery" and "queuing" (Anderson, 2003, pg. 26).

Sophisticated error recovery is the process of attempting to re-read a particular sector. This is achieved by moving the head slightly, changing the timing and repeating the process a set number of times. This in turn means that there is a much better chance of reading all the data in a sector (Anderson, 2003, pg. 26).

Queuing is very straightforward to understand. Imagine making a shopping list where the person just writes down what they think they might need such as “bread”, “milk”, “pears”, “croissants”, “dish washing liquid”, “apples”, “cream” and “bananas”. In a non-queued system the person starts at the supermarket entrance and goes through the list one item at a time, which could mean retracing their steps repeatedly and taking longer to collect the items on the list.

A more observant person would notice that “pears”, “apples” and “bananas” are all likely to be in the “fruit” section, and queuing means the person goes to the “fruit” section and the three items are collected together. Likewise, “bread” and “croissants” are likely to be in the same or similar section of the supermarket so collecting those items at the same time makes sense. For hard drives this works in a similar method with “reads” queued in such a way as to make more efficient use of the timing or head position.

2.2.2.1 Small Computer System Interface (SCSI)

For those with an interest in the SCSI interface, the technical committee for SCSI storage systems can be found at www.t10.org (last accessed 23/11/2011).

The history of SCSI started in 1979 with Shugart Associates Systems Interface (SASI).

> “The goal was to develop a drive interface that supported logical addressing of data blocks instead of physically addressing of cylinders, heads and sectors. … Such an interface would end the compatibility problems associated with bringing new drive technologies to market…The new interface would allow computer manufacturers to develop hard disk drivers that were able to recognize the properties of the connected disk drives themselves” (Schmidt, 1997, pg. 87).

IBM (2011) host a web page presents twelve different connectors for SCSI drives, predominantly the male and female pin equivalents of the 50 pin, and 68 pin variants of SCSI drive. This page also presents a comparison between different SCSI modes, which include seven variants of regular, fast, wide, fast-wide and ultra. Those differences include bandwidth (8-bit or 16-bit), max transfer speed (ranging from 5 MB/second for SCSI mode to 80

MB/second for Fast/Wide Ultra2 SCSI) and the maximum cable length (without using repeaters) which ranges from 1.5 metres to 12 metres. Ultra SCSI-320 as the name implies has a maximum transfer rate of 320 MB/second (Kumar et al., 2011, pg. 1675).

There is also SCSI-SCA "hot swap", which uses an 80 pin connector. One of the differences between SCSI and SCSI-SCA is that SCSI hard drives have two plugs, one for the data cable and one for the power supply, whereas the SCSI-SCA connector combines both into one plug. Hot swapping is the ability to be able to physically remove a hard drive from the system without having to reboot the system (Dufrasne et al., 2004, pg. 3). This feature is practical for systems that should have limited or no downtime such as enterprise servers, commercial websites and the like.

The final feature of note regarding SCSI drives is that a SCSI cable can support either 8 or 16 devices (such as CD-ROM drives, hard drives, external scanners and the like), depending on the version used. In turn the SCSI drive does not have external jumpers but the interface itself handles the setting of the "SCSI_ID" so the host knows which physical device is which.

2.2.2.2 Integrated Device Electronics (IDE) aka ATA

For those with an interest in the IDE/ATA interface, the technical committee for AT Attachment can be found at www.t13.org (last accessed 23/11/2011).

Compared with SCSI, IDE typically uses a standard 40 pin cable (Dufrasne, et al., 2004, pg. 3) that supports only two devices, a "master" drive (drive 0) and a "slave" drive (drive 1) (Schmidt, 1997, pp. 33-34), whereas SCSI supports 8 or more devices.

The settings for master and slave are typically controlled by the use of jumper settings (as shown in Figure 1 on page 6). These are a small group of separate pins (not to be confused with cable pins) that can be set using a plastic encased electro-conductive "jumper". Depending on the location and orientation of the jumper it informs the interface whether the drive is the master or the slave.

Dufrasne, et al. present the following as an overview of IDE (2004, pg. 2):

> "Although there have been several developments to the specification that added performance (the original speed was just 3 MBps when the protocol was introduced) and reliability features ... and multiple data-transfer modes, including Programmed Input/Output (PIO), direct memory access (DMA), and Ultra DMA (UDMA), it has changed little in the past 10 years. Its design limitations, coupled with

faster applications and PC processor performance, meant that it was often the cause of bottlenecks in data transfer, because it has achieved its maximum data transfer rate of 133 MBps."

2.2.2.3 Comparison between SCSI and IDE

Anderson et al. (2003, pg. 245) present an argument "that SCSI and ATA drives are the same technology internally but differ only in their external interface and in their suggested retail price". They referred to "Enterprise Storage" (ES) and Personal Storage (PS) when developing their argument. Enterprise storage was described as typically being attached to very large systems that store large quantities of data and that were expected to support many users simultaneously, potentially in a 24/7 environment. Personal storage, in contrast, was used for a few hours a day, on a single user system and therefore the power requirements and data access requirements were less (Anderson, et al., 2003, pg. 245).

They go on to state that the ATA interface was a more simple interface (due in part to only having to deal with the master and slave relationship rather than having to interact with more devices) and that in part led to the ATA interface being low cost whereas SCSI drives must be optimized for performance, reliability and the ability to connect to multiple hosts (Anderson, et al., 2003, pg. 255).

In terms of data remanence research Anderson et al. (2003) are arguing that attempting to target companies versus individuals based solely on the interface type would not be an effective strategy. However Jones (2009, pg. 3) states that companies can be effectively targeting by purchasing SCSI drives and Valli and Woodward's research (2008) used a hard drive selection bias of selecting only SCSI drives for company profiling.

From the New Zealand perspective there have been a number of computer surveys such as Quinn (2010) and Johnson, et al. (2009) however none of those surveys have asked specifically about the interfaces used by companies. That is likely because it may be deemed too obscure or too irrelevant for those researchers' purposes.

Quinn (2010) surveyed New Zealand businesses but only asked the type of operating system used rather than interface. Johnson, et al. (2009) considered computer hardware from the desktop and server perspective. The direct question of interface types was not asked; the closest question related to the vendor of the hardware that schools purchased. For desktop

machines popular answers were well known companies such as “Compaq” and “Dell” (Johnson et al., 2009, pg. 44) and it is probable that those machines used the cheaper IDE drives. Under the “Servers” results, “Compaq” and “IBM” (Johnson, et al., 2009, pg. 45) led the list but there is not enough information to determine if the servers would have IDE or SCSI drives.

2.2.2.4 Serial AT Attachment (SATA)

As the name implies, Serial ATA is part of the ATA family and therefore is overseen by the T13 committee, but the Serial ATA working group located at http://www.serialata.org (Dufrasne, et al., 2004). The intent of SATA was to introduce a serial system, as ATA was typically parallel.

One of the features and improvements of SATA compared with IDE that Dufrasne, et al. (2004) focus on is that the maximum date rate for SATA 1.0 was 150 MBps which is at least 10% greater than the previous ATA maximum of 135MBps. Another benefit is the pin count decrease from 40 pins to 7. Dufrasne et al. (2004, pg. 3) explains the benefit of decreased pin count as:

> “Parallel ATA uses 16 separate wires to send 16-bits of data and thus must use a bulky flat cable, which is the cause of electromagnetic interference that compromises data integrity.”

A third benefit of SATA when compared with IDE is that SATA does not use jumpers, as it only allows one device per cable. That in turn gives an added benefit to data remanence research as it means one less point of failure (missing, damaged or incorrectly configured jumpers). Krutov (2011) describes SATA II and SATA III. The following is a list of selected features (Krutov, 2011, pp. 2-3):

> “SATA storage interfaces have the following characteristics:
>
> - Interface speeds of ... 150 MBps, 300 MBps, and 600 MBps
> - Single hard drive capacities of 250 GB, 500 GB, 1 TB, 2 TB, and 3 TB
> - Support for Self-Monitoring, Analysis, and Reporting Technology (S.M.A.R.T.)”
> - Support for Native Command Queuing (NCQ), which allows you to reorder disk data access command sequence, optimizing seek time by minimizing physical movement of magnetic heads over disk plate”

From a data remanence perspective, the SATA standard includes the SECURE ERASE functionality which means those drives can be easily sanitised before disposal.

2.2.2.5 Universal Serial Bus (USB)

As the name implies, USB is a bus architecture. USB can support up to 127 devices including external devices such as scanners, keyboards and mice (Axelson, 2001).

Axelson (2001, pp 1-7) further describes the benefits of USB as including "fast", "reliable", "flexible" (many different devices can use a USB interface), "inexpensive" and "power-conserving".

USB storage devices can be classified into two main types. The first type is the large capacity hard drive, which are typically the same physical size as a traditional internal hard drive, typically in the 100 Gigabyte - 2 Terabyte range. These hard drives can either use the computer itself as the power source, or can use mains supplied power. The second type of USB storage device is the "pen drive" or "thumb drive". As the name implies these drives are very much smaller in physical size and capacity, typically being the same size as a human thumb. They use the computer as their power supply only, making the drives very portable, but also very easy to misplace and damage. Some brands of USB thumb drive have added security features such as encryption which can mitigate problems if the thumb drive is lost and subsequently found by someone else.

With regards to data remanence there are two main issues. The first issue is that USB storage devices are typically external which means they can easily be transferred from one computer to another and are therefore also good for storing files rather than operating systems.

The second issue is the limited research on USB storage devices. Jones et al. (2009) and Chaerani et al. (2011) have conducted research into USB thumb drives but there does not appear to be research conducted explicitly on the larger external hard drive style USB storage devices. One reason is people may keep their external hard drives rather than resell them. Much like the issue with "solid-state" drives on the second hand market, searching the Trade Me auction site for USB hard drives returned three main results. The first was "enclosures" (converters that have an IDE plug and a USB plug so the IDE drive can become external), the second was brand new USB external hard drives and the third was brand new USB thumb drives. There were practically no lots on sale that were clearly second hand USB storage devices.

Sourcing USB thumb drives from "lost and found" type services was considered a possibility for this research, however as users had lost those thumb drives it would be expected that all the data would be remaining on the drives (other than drives that users had encrypted) and it would not accurately reflect data remanence.

2.2.2.6 Adapters

As the name implies, adapters provide a mechanism to convert one connector type into a different connector type. A good example of this might be using an IDE to USB connector to be able to use an internal IDE drive as an external USB drive.

2.2.3 Interactions between Hard Drives and Operating systems

There are a number of considerations for data remanence regarding the interactions between the hard drive and the operating system. While a number of those interactions have been abstracted into the interfaces, there are things such as hidden disk areas, master boot records and partitions that are relevant to data remanence.

2.2.3.1 Hidden Disk Areas

There are two main areas that allow hard drive manufacturers and others to hide data on a hard drive. Those are the Host Protected Area (HPA) and the Device Configuration Overlay (DCO). The HPA "is designed to store information in such a way that it cannot be easily modified, changed, or accessed by the user, BIOS, or the OS" (Gupta et al., 2006, pg. 1). From a data remanence perspective it is important to know if the hard drive has a HPA and if the tools that are being used to capture/analyse the hard drive can find the HPA, since an advanced user may hide data there.

The DCO is a feature introduced in the sixth version of the ATA standard, and is meant to enable a consistent number of sectors on a hard drive that have slightly different physical capacities (Gupta, et al., 2006, pg. 4). This feature is relevant to data remanence research as it could indicate a difference between the stated capacity size and the actual capacity of the drive.

One of the reasons given for having hidden areas on disk drives is to allow computer vendors to put "system restores" or other system rebuilding tools in a "safe" place that normally would not be altered. This in turn makes it easier than having to supply DVDs that may get lost or damaged if a user wishes to rebuild their system.

2.2.3.2 Master Boot Record

The Master Boot Record (MBR) and its role in computer start-up is described in a paper by Mohamad and Mat Deris (2011). The Master Boot Record is deemed to be a "consistent starting point" for the hard drive, and contains important information such as how many partitions the drive has, and what sort of partitions they are (primary or extended). To make the MBR consistent, it is located at cylinder 0, head 0, sector 1, the first sector on the disk (Mohamad & Mat Deris, 2011, pp. 1-2).

Depending on the software tools used for analysing hard drives, where the data have been deleted it may be possible to find the MBR and determine useful information from it. That information can also include if and how the hard drive had been wiped (for example, Figure 7 on page 83). Hard drives that use GUID Partitioning instead of MBR partitioning (explained in the next section) may instead have a "Protective MBR" which allows for backward compatibility with disk management utilities that operate on MBR (MSDN Microsoft, 2011).

2.2.3.3 Partitions

Partitions are described by Microsoft (2011) as:

> "A partition is a contiguous space of storage on a physical or logical disk that functions as if it were a physically separate disk. Partitions are visible to the system firmware and the installed operating systems. Access to a partition is controlled by the system firmware before the system boots the operating system, and then by the operating system after it is started."

There are two main types of partitioning when using Microsoft products, the MBR and the GUID Partition Table (GPT). As explained by Microsoft, MBR partitions have some issues such as (MSDN Microsoft, 2011):

> "MBR disks support only four partition table entries. If more partitions are wanted, a secondary structure known as an extended partition is necessary. Extended partitions can then be subdivided into one or more logical disks.
>
> By convention, Windows creates MBR disk partitions and logical drives on cylinder boundaries based on the reported geometry, although this information no longer has any relationship to the physical characteristics of the hardware (disk driver or RAID controller).
>
> MBR partitioning rules are complex and poorly specified."

Alternatively GPT partitions have a variety of improvements such as (MSDN Microsoft, 2011):

> "GPT provides a more flexible mechanism for partitioning disks than the older Master Boot Record (MBR) partitioning scheme that was common to PCs.
>
> The GPT disk partition format is well defined and fully self-identifying. GPT disks use primary and backup partition tables for redundancy and CRC32 fields for improved partition data structure integrity. The GPT partition format uses version number and size fields for future expansion.
>
> Each GPT partition has a unique identification GUID and a partition content type, so no coordination is necessary to prevent partition identifier collision. Each GPT partition has a 36-character Unicode name. This means that any software can present a human-readable name for the partition without any additional understanding of the partition."

This leads to the benefits of partitions overall such as the ability to have more than one operating system on a computer such as Windows and Linux, and there may be some efficiencies for data storage/defragmentation as well. Another reason partitions are used is to have operating system files in one location, and software or data stored in another location. The advantage of this is that if a user wishes to back up their data, they know they only need to back-up one specific partition. From a data remanence perspective it also means some users may only properly wipe their "data" partition and leave the operating system partition alone. It should be noted that there may be user related data (user names, and other settings) remaining in operating system files.

2.2.4 Other data storage devices: Mobile/Smart Phones

Williamson et al. (2006) are among the first to consider mobile phone data remanence and outlines a number of issues surrounding the attempts to forensically analyse mobile phones. One of the first challenges Williamson et al. faced was that there was a variety of different "handsets" for cell phones which in turn have a large number of different features. Those features include the possibility that the phone may have "Bluetooth" or "infrared" connectivity, whether the phone is capable of storing SMS messages or not, and how much memory was stored onboard, on SIM cards or on removable memory cards (Williamson, et al., 2006, pg. 125). A related challenge was the variety of different data transfer cables that cell phone manufacturers have created; ultimately Williamson et al. could only examine two

specific models of cell phone due to not being able to acquire the right sort of cables (Williamson, et al., 2006, pg. 127).

Additionally there is the challenge of the different types of data that a cell phone might store, ranging from recorded audio messages or ringtones, computer files or images, calendaring and specific settings which are not consistent across all brands and models of cell phone.

Initially Williamson et al. selected four different software tools: *TULP 2G*, *Mobiledit! Forensic*, *Cell Seizure* and *Oxygen Phone Manager*. However, *TULP 2G* was not able to be used, and *Oxygen Phone Manager* was "used only sparingly". They found that the functions performed and functions promised by the software tool vendors came out as either "unknown", "non applicable" or "not found" which suggests a number of potential issues with the tools.

One of the bigger concerns they had were the hashing functions used, MD5 and SHA-1. *Cell Seizure* was the only tool where Williamson et al. could find where hashing functions worked, and they note that that functionality was unreliable (Williamson, et al., 2006, pg. 131).

For example they observed that certain phones (different models) being tested returned the same hash value, and that only one image appeared to be stored on the phone. They theorised that the hash function was not hashing the entire phone or the destination copy of the phone but rather hashing just the logo image. This raises the concern that either only the logo image was being copied or that the software's mechanism for calculating hash functions is not overly useful for research purposes.

They conducted further research with the phones and determined there were potentially four different MD5 hash values for three different tests. As they observed this could cause problems if the results were being presented in a court of law and felt overall the MD5 hashing function in *Cell Seizure* was inadequate.

Another challenge was that at the time "Extracting deleted data from the handsets was impossible" (Williamson, et al., 2006, pg. 132). However it was noted that it may be possible to forensically capture the memory cards from the phones (if applicable) and then examine those captures using tools such as *Encase*. This assumed that the phone model supported those types of memory card and that the software would be able to do that.

The other issue that Williamson et al. had was network connectivity. The phone may have functionality (such as a clock update feature) that means it is constantly receiving data or updates when it is switched on. That in turn means a full forensic capture would have hashing problems since the source may have changed during the capture process. Furthermore the stored data on the phone such as call logs, images and SMS messages may have changes during the capture process which could also present complications.

Finally Williamson et al. concluded their research was a partial success. They observed that they could not retrieve deleted data from phones and that funding constraints partially limited their research to older handsets. As noted, one of the bigger challenges for cell phone researchers is that there can be many different cable types required as manufacturers are not consistent with their cabling.

Many of the issues Williamson et al. raised were still relevant at the time the experiments for this thesis were designed; Owen et al. (2010) and Grispos et al. (2011) also faced similar challenges. Cabling, forensic software maturity and the myriad of cell phone brands, models and operating systems are all potential issues that were considered. Hard drives, on the other hand, only require a few standardised cables (even with SCSI and the variants for that), the tools such as *Encase* and *Forensic Tool Kit* have been rigorously tested and the ability to use a dead forensics methodology means that data are much less likely to be altered during the capture process.

Owen et al. (2010) present a number of challenges to conducting data remanence research on mobile phones such as "new handsets being introduced every 4 days, most software will only produce a report on a limited section of the mobile phone data and the interpretation or reconstruction of deleted data is ad-hoc, incomplete and difficult to read" (Owen et al., 2010, pg. 26). These challenges are very similar to the challenges that Williamson et al. (2006) faced.

Grispos et al. (2011) also conducted research into mobile phone forensics with a specific focus on Windows operating system smart phones. Even when limiting themselves to a subset of the mobile phone market, Grispos et al. still found a number of challenges with forensic investigation for mobile phones, notably "that no one technique recovers all information of potential forensic interest from a Windows Mobile device; and that in some cases the information recovered is conflicting." (Grispos, et al., 2011, pg. 23)

Grispos et al. outline the two main ways of acquiring data from a smart phone: "physical" and "logical". Typically physical acquisition requires either physically de-soldering the memory chips from the phone, which could permanently damage the phone, or by using specialised hardware such as the *Cellebrite's Universal Forensic Extraction Device* (*CUFED*) to connect via the *JTAG* (Joint Test Action Group) ports that may not be accessible on all models of cell phone. (Grispos, et al., 2011, pg. 24). *JTAG* ports contain "circuitry that may be built into an integrated circuit to assist in the test, maintenance, and support of assembled printed circuit boards. The circuitry includes a standard interface through which instructions and test data are communicated." (IEEE, 1990). Physical acquisition is akin to microscopy techniques for hard drives (explained in the next section) and therefore is considered outside the scope for this research. Logical acquisition uses software to recover "logical objects" from the mobile phone (Grispos, et al., 2011, pg. 24). "Logical objects" in this case typically covers known file types such as PowerPoint and JPEG. As Grispos et al. and other researchers into mobile phones have stated the software used may not be able to find all the data currently and previously stored on a mobile phone. In other words using logical acquisition is easier and less invasive but does not guarantee finding all the data. They conducted a set of experiments with a cell phone with specific known data on it to compare various different data carving tools to determine whether the tools could find all the deleted data or not. As an explanation of data carving (aka file carving) , Garfinkel (2007) defines carving as (2007, S2):

> ''File carving' reconstructs files based on their content, rather than using metadata that points to the content. ... file carving can recover files that have been deleted and have had their directory entries reallocated to other files, but for which the data sectors themselves have not yet been overwritten."

Grispos et al. noted that in certain circumstances the toolkits often used for hard drive forensic analysis (*Encase* and *Forensic Tool Kit (FTK)*) were not able to analyse the forensic images from their experiments (Grispos, et al., 2011, pg. 35).

As those researchers have shown, specialised tools and techniques are required to conduct research on data storage devices other than hard drives, and those tools and techniques are either not always reliable or there are special challenges regarding some storage devices that need to be considered before the experiments can be designed.

2.3 Technical issues regarding Hard Drive Data Remanence

This section considers research that involve technical issues surrounding data remanence for hard drives and is split into that focused on hardware and that focused on software.

One of the most commonly referenced articles in understanding the technical issues surrounding secure data deletion is by Gutmann (1996). That paper introduces the "Gutmann method" as a method for overwriting data in such a way that it is unlikely to be recoverable. Gutmann outlines how data are written to hard drives, and how those data can be retrieved.

As Gutmann's abstract states (1996, pg. 1):

> **"**This paper represents an attempt to analyse the problems inherent in trying to erase data from magnetic disk media and random-access memory without access to specialised equipment, and suggests methods for ensuring that the recovery of data from these media can be made as difficult as possible for an attacker."

Gutmann first considers technologies that were available in 1996 such as "Magnetic force microscopy (MFM)", "scanning probe microscopy (SPM)" and "Magnetic force scanning tunnelling microscopy (STM)" and describes the processes used.

Microscopy is explained by first describing how data are written onto a disk. While hard drives use a simple binary ("1" or "0") encoding system for data Gutmann states that if a 0 is overwritten with a 1 the value can be interpreted as being 0.95, and that if a 1 is overwritten by a 1 the value is 1.05. While the computer ("normal disk circuitry" is the term Gutmann uses) reads both 0.95 and 1.05 as "1", specialised circuitry can determine more accurately if the value is 0.95 or 1.05 and therefore what the previous layer was. Gutmann did not state the effect of overwriting 1 with 0, so this is ignored in the following example.

As a simple example if "1.05, 0, 0.95, 1.05" was found, a normal computer would read it as "1,0,1,1". By using microscopy as described by Gutmann, it would both show "1,0,1,1" but also show that the underlying value of "1,0,0,1" had been overwritten.

Gutmann states that it is possible with the right equipment, time and training that physical images of hard drive platters could be captured and analysed, giving reasonable indications of what data may have been left behind from at least two or three "generations" (or deletions)

before. At the time of that writing (1996) platters could be captured in a matter of minutes, and individual tracks found with some skill in approximately 10 minutes. It is unclear the time it would take on modern hard drives which would typically have a much larger capacity than the drives Gutmann would have examined.

Central also to Gutmann's method is understanding the various encoding systems used for writing data to media. Encoding systems are used much like full stops are used in sentences. It allows the system to understand the size and nature of a block of data that has been read which can also become useful when considering data carving.

Gutmann also considers issues surrounding "run-length encoding" and why it may make a difference to target media. From the explanation of the run length encoding and other schemes this leads to deriving "The Gutmann method". That method uses a set of 22 overwrite patterns and recommends 35 consecutive writes in total to "delete" the data to the point that microscopy and other techniques cannot be used to reconstitute it (Gutmann, 1996, pg. 8). One of the issues not addressed by Gutmann is the likely time taken to run 35 passes of the overwriting function on hard drives which in turn could turn a number of people away from using the technique to remove data.

Since Gutmann's work was published 15 years ago there has been a reasonable amount of debate (Jouvok et al., 2006; Wright et al., 2008), including what constitutes an appropriate number of passes, whether this method is appropriate for intelligence agency wiping or techniques such as degaussing be used, and whether the method is still worth using due to "SECURE ERASE" which is part of the ATA specification. The National Institute of Standards and Technology (NIST) states that ATA disk drives manufactured after 2001 (over 15 GB) can effectively be cleared and purged by one overwrite (NIST, 2006, pg. 20).

A different consideration for hardware data remanence is the data storage device is faulty, damaged or otherwise unreadable. Sutherland et al. (2010) consider that repairing unreadable hard drives may yield identifying information. This is based on the three reasons why they believe second hand hard drives may be faulty on arrival: first the drives may have been damaged in transit; second the seller knew the drive was faulty when sold; or third the seller had not used the hard drive for a while and it subsequently was not working but the seller

believed it was still working (Sutherland, et al., 2010, pg. 159). They also note that in the first and third cases the seller may not have had the time to remove the data.

Sutherland et al. (2010) describe the process of classifying, repairing and then attempting to forensically capture those unreadable drives. They list five major categories of hard drive fault with examples and reasons provided for why the drive may be faulty. They then consider what repairs could be made and the benefit from making those repairs, which is then used to present a four point taxonomy of possible repairs. The taxonomy ranges from "a simple fix" that does not require specialist tools or donor parts, repairs that require either specialised tools or donor parts, to "no repairs possible" (Sutherland, et al., 2010, pg. 160).

Other research into hardware and software for data remanence has been carried out by National Institute of Standards and Technology (NIST) which is an American government agency that conducts independent research. They created the *Computer Forensics Tool Testing* (*CFTT*) program, which as the name suggests, tests a variety of computer forensics tools and provides detailed reports on the reliability of those tools under specific diverse testing situations.

Lyle (2012) is a handbook that outlines approximately 100 hardware devices or software programs that have been tested for specific forensic functionality such as *Encase 3.20*, *FTK Imager 2.5.3.14* and *Tableau Imager (TIM) Version 1.11* under the "Disk Imaging" category; *Tableau Forensic Duplicator Model TD1 (Firmware Version 3.10)* and *Voom HardCopy II (Model XLHCPL-2PD Version 1.11)* under the "Forensic Media Preparation" category; and *CelleBrite UFED 1.1.3.3*, *MOBILedit! Forensics 3.2.0.738* and *XRY 5.0.2* under the "Mobile devices" category.

The handbook then provides a brief (one or two page) summary of each tool and how it performed under test conditions, and then provides a link to the specific report (often 30-60 pages) on the CFTT website. An example is the *Encase 6.5* report that stated:

> "For some removable USB devices (Flash card and thumb drive) that have been physically acquired, there may be a small number of differences in file system metadata between the image file and the restored device (DA–14–CF and DA–14–THUMB)." (Lyle, 2012, pp. 22-23)

The full report outlines the testing parameters, and any anomalies that may have been found. One of the key benefits of the handbook is that it presents a new investigator or researcher with a long list of suitable, tested tools, and situations that the investigator or researcher may need to consider when using those tools.

2.4 Forensic Computing

This section addresses the concepts of forensic computing including technical and legal aspects.

2.4.1 What is Forensic Computing

A straightforward definition of Forensic Computing consisting of four elements is:

> "[T]he process of identifying, preserving, analysing and presenting digital evidence in a manner that is legally acceptable" (McKemmish, 1999, pg. 1)

A number of authors such as Sherman (2006) however point out that evidence generated from forensic computing also needs to be able to be understood "by the people that need to understand it, the judiciary and the jury." (Sherman, 2006, pg. 135)

McKemmish (1999, pp. 3-4) also then presents four rules for forensic computing relevant to presenting evidence in a court of law. Those rules are:

- "Minimal handling of the Original"
- "Account for any change"
- "Comply with the rules of evidence"
- "Do not exceed your knowledge"

The rest of this section will consider two elements of McKemmish's definition of forensic computing (preserving, and analysing digital evidence) and three of McKemmish's four rules as they apply to data remanence research. McKemmish's rule of "Do not exceed your knowledge" is considered self explanatory.

2.4.2 Minimal handling of the Original

One of the easiest ways of ensuring there is minimal handling of the original evidence source is to forensically clone or capture the original source and then analyse the copy. Depending on the original data source there can be technical issues surrounding exactly how to make a

clone of the source. As such there are two main different methods “Dead forensics” and “Live forensics” for capturing the data source.

2.4.2.1 Dead forensics

Dead forensics involves powering off the source device and then capturing the data. In the case of hard drives this may involve connecting the source device to a forensics capturing device such as the *Tableau TD-1* (referred to as *Tableau* in this thesis). The *Tableau* is further described in Appendix G. The capturing device then supplies power to the hard drive to enable data to be read off of it, but also provides mechanisms to ensure the hard drive’s data are not altered in the process. For the capture device to be useful for data remanence and forensics investigation it should perform a bitwise copy reading from the very first bit (either a “0” or “1”) on the source device to the very last bit and copy all the bits that it finds. Additionally it should have a mechanism such as hashing (described in Section 2.4.3 on pages 27 and 28) in place to ensure that the data have been copied exactly and not altered during the process.

The key benefit of dead forensics is that once the storage device has been powered off, it can be captured in a timeframe that suits the investigator. In the case of hard drives once the drive is turned off, the data are unlikely to change and it makes for a more accurate image of what is on the hard drive. It should be noted that the process of shutting off the hard drive may alter the data and there is debate in forensics investigation circles as to whether the hard drive should be properly shut down, or whether the power cord should be pulled out to force the hard drive to shut off.

The downside to dead forensics is that some systems either cannot or should not be shut down before capturing is started. The first situation is where the source device has used encryption. In this situation the source device data are expected to be encrypted (not easily readable) when it is turned off. Typically when the source device is turned on, the user enters a password or phrase and the data are then unencrypted and easily readable. Therefore it makes sense to try and capture the data when the source device is turned on (or acquire the encryption password from the user).

Adelstein (2006) describes a number of other factors or reasons why dead forensics may be difficult in a number of common situations. Those factors include the size of the drives being

examined, that servers may be being used for e-commerce applications and therefore companies may lose business if the entire system has to be shut down and that the context of the information captured may have an effect on the subsequent analysis (Adelstein, 2006, pg. 64). When these problems occur a solution is to use "live forensics".

2.4.2.2 Live forensics

Live forensics can be described as

> "Traditional digital forensics attempts to preserve *all* (disk) evidence in an unchanging state, while live digital forensic techniques seek to take a snapshot of the state of the computer, similar to a photograph of the scene of the crime."
> (Adelstein, 2006, pg. 64)

A number of researchers (Williamson et al., 2006; Owen et al., 2010; Grispos et al. 2011) point out the challenges of conducting live forensics on mobile phones and their data storage devices. These challenges typically involve determining if the device's data are changed during the capture process (from incoming messages, specific data from the cellular network, clock updates and the like).

2.4.3 Accounting for Change

This section first considers the "chain of custody" which is a mechanism to account for the location of and potential accesses to the evidence at any time from initial contact by the investigator until it reaches court. This section then considers the technical aspects of determining if electronic evidence has been changed.

Giannelli explains the chain of custody as (Giannelli, 1983, pg. 534):

> "In addition to showing that the object introduced in evidence is the same object as the one involved in the crime, the proponent of evidence must show that the object has retained its relevant evidentiary characteristics. Substantial alteration of the item reduces or negates its probative value and may mislead the jury"

Typical aspects of the chain of custody are documenting who had control over the evidence, when the evidence was handed to someone else, how and where the evidence was stored and showing how secure that location was. For example demonstrating that only people with electronic swipe cards could access the building that stores the evidence means it is a lot less likely someone else has altered the evidence.

A mechanism to determine that there has not been change to the data is hashing. In computer forensics hashing can have at least two slightly different but related functions, both of which were used in this research.

The first function relates to the data acquisition phase and is to ensure that the source hard drive and the target acquired image are identical. This is achieved by using a "hash function" to mathematically compute a single value for the source drive and a single value for the destination drive after the source drive has been captured.

As an example with very simplified hash values, the source drive might be just be the word "cat" and hash to 3120. If the destination drive also hashes to 3120 then it is assumed that the destination drive has "cat" on it. If during the capturing process the destination drive ends up with "bat" written on it the hash value for that might be 2120 and therefore this shows that the destination drive does not match the source drive. The actual determination of the hash values, and how it can be assumed that if the source drive and destination drive hash values match then they are identical is explained by a number of authors, however Thompson (2005) is one of the easier to understand.

Thompson (2005, pg. 36) outlines the "three fundamental properties" of hash functions in this context as:

> "[1] A hash function must be able to easily convert digital information (i.e. a message) into a fixed length hash value.
> [2] It must be computationally infeasible to derive any information about the input message from just the hash.
> [3] It must be computationally infeasible to find two files that have the same hash. Hash(Message1) = Hash(Message 2)."

The first two points are reasonably straightforward, but the third point can be technically challenging. Thompson explains it by using the example of the "Birthday Paradox". The "Birthday Paradox" considers two elements of people's birthdays. The first element is "How many people must enter a room such that there is a 50% chance someone else shares <u>your</u> birthday?" It requires on average 183 people to meet this criteria, since it is matching one particular pair of dates (such as February 11). The second element is "How many people must enter a room such that there is a 50% chance <u>any two people</u> share the same birthday?"

The answer is only 23 people as that generates a total of 253 different pairs of dates. This would seem to be a relatively small number of people (such as children in a classroom, a departmental staff meeting or a two sports teams competing for example). This is Thompson's point: that a hash function needs to ensure that the hash values it generates for files will not accidently "collide" with other files. Thompson (2005, pg. 39) then considers this in forensic computing terms:

> "In the real world the number of files required for there to be a 50% probability for an MD5 collision to exist is still 2^{64} or 1.8×10^{19}. The chance of an MD5 hash collision to exist in a computer case with 10 million files is still astronomically low."

The second function of hashing relates to the analysis phase. This form of hash function allows known files to be filtered from the investigation or examination if they are contained in a library of known files. The "National Software Reference Library" (NSRL) is a project that hashes known files to create a library suitable for filtering. Mead (2006) states that a forensic examiner searching a Microsoft Windows 2000 system for images may find 4000 images are part of the typical installation. Using a technique that quickly recognises those files as being part of the installation then they can be ignored and time can be more efficiently used examining the remaining files (Mead, 2006, pg. 139). Law enforcement can also use a specific library to filter known objectionable, illicit or pornographic images to find them.

2.4.4 Comply with the rules of evidence

McKemmish (1999) only briefly touches on the meaning of "comply with the rules of evidence" but the salient point is that "the presentation of evidence should not alter the meaning of the evidence". Dicarlo (2001) gives an in depth review of the "rules of evidence". He states "There are four traditional types of evidence: real, demonstrative, documentary, and testimonial." and then describes the general rules of admissibility and includes "relevant, material, and competent" (Dicarlo, 2001). With regards to data remanence, competency can be measured by determining if the tools used are suitable or have been NIST evaluated and if the investigator is qualified to undertake the investigation. As such certain forensic tool vendors such as Guidance Software (makers of *Encase*) thus offer certification programmes.

2.4.5 Implications for this research

Forensic computing has a number of requirements for law enforcement investigators to ensure that evidence they present in court is legally acceptable, and understandable. Although it was considered unlikely that material would be found in the experiments for this thesis that would require law enforcement intervention, the methodology for this research ensured that the experiments followed forensic computing requirements. This was achieved by using a dead forensics approach, ensuring the hard drives were stored as securely as practicable (with no outside interference) and by using hashing to ensure the hard drive captures were identical to the source drives. The tools used were also recognised as being forensically sound and fit for purpose.

2.5 Psychology of Security

There can be technical reasons and emotional or psychological reasons why people do not properly sanitise storage devices before disposing or selling them. This section considers the overall psychology of security then applies those considerations to data remanence issues in Section 2.5.1. Two authors Schneier (2008) and West (2008) consider a number of aspects of the psychology of security and review concepts and experiments in the related fields.

Factors and theories that explain the psychology of security include:

- The concept that security is a trade off
- How people evaluate risk
- Prospect theory
- Optimism bias
- Safety is an abstract concept.
- Feedback and learning from security-related decisions.

With regards to the concept that security is a trade off Schneier (2008) uses the example that to prevent further terrorist attacks by airplanes such as the “9/11” attacks, a solution would be to ground all planes (Schneier, 2008, pg. 2). While this would drop the risk/threat of airplane based terror attacks to practically zero it would be impractical, and so there is a trade off between a very unlikely but very devastating event happening and by having a very useful service such as air travel.

This in turn introduces the considerations for how people evaluate risk. From a human perspective there are a number of reasons why people are bad at evaluating risks that include (Schneier, 2008, pg. 4):

- People exaggerate spectacular but rare risks and downplay common risks.

- People have trouble estimating risks for anything not exactly like their normal situation.

- Personified risks are perceived to be greater than anonymous risks.

- People underestimate risks they willingly take and overestimate risks in situations they can't control.

- Last, people overestimate risks that are being talked about and remain an object of public scrutiny.

Clearly there is a problem if people have difficulty in estimating risk and therefore how they can make effective trade offs in those situations. To better understand how people make decisions Schneier considers Utility theory and Prospect theory using examples of the work of Kahneman and Tversky who developed Prospect theory.

Kahneman and Tversky (1979) initially consider "expected utility theory" as "the utilities of outcomes are weighted by their probabilities" (Kahneman & Tversky, 1979, pg. 265). They further clarify that utility theory relies on three tenets "expectation", "asset integration", "risk aversion" (Kahneman & Tversky, 1979, pp. 263-264). A simple example would be the problem of "Would you prefer $1 for free, or would you prefer to flip a fair coin, if it lands heads collect $2, if it lands tails, collect $0?"

They observed a number of effects such as certainty effect "We first show that people overweight outcomes that are considered certain, relative to outcomes which are merely probable" (Kahneman & Tversky, 1979, pg. 265) and the reflection effect (Kahneman & Tversky, 1979, pg. 268) which considers that when the participants are queried about negative outcomes as opposed to positive outcomes such as "Either lose $1, or flip a coin to either lose $2 or lose $0". Kahneman and Tversky also consider a number of variables

including the different value of the outcomes (flipping a coin for $60 000 is not the same as flipping it for $2).

Prospect theory is based on the observation that when presented with a problem pair that people do not pick the same answer for both problems. Schneier presents a much clearer example than Kahneman and Tversky do as they consider many different problem pairs and consider different variables for their problem pairs.

The first problem is "[A] A sure gain of $500" and the second option is "[B] A 50% chance of gaining $1000 and a 50% chance of gaining $0". The second problem is "[C] A sure loss of $500" and "[D] A 50% loss of $1000 and a 50% chance of a $0 loss".

"Utility Theory" suggests that people would pick A and C in the same ratio that they picked B and D. That is those who selected a sure gain of $500 would also select a sure loss of $500. The results from that specific experiment were that people picked [A] 84% of the time and if "Utility Theory" was applicable then [C] should have also been 84%. However, [C] was only picked 30% of the time and this leads to "Prospect theory" (Schneier, 2008, pg. 9)

By itself Prospect Theory explains a lot of computer security, that while people may not be good at estimating the risk of something, they prefer to take the risk of something very bad happening or nothing bad happening, as opposed to something bad definitely happening to them. However there is another factor that explains why people make these decisions which is "Optimism Bias".

"Optimism Bias" is described in studies by Weinstein (1980). Weinstein noted that "It is usually impossible to demonstrate that an individual's optimistic expectations about the future are unrealistic" (Weinstein, 1980, pg. 806). This is because an individual may actually be correct or that it may be difficult to measure the expectation easily. However Weinstein considered that testing groups of people would be able to determine "if all people claim their changes of experiencing a negative event are less than average, they are clearly making a systematic error, thus demonstrating unrealistic optimism" (Weinstein, 1980, pg. 806).

Weinstein's experiments involved asking college students to consider 42 events which were either negative or positive, such as getting a good job after graduation or developing a drinking problem and then rate themselves versus their fellow students. Weinstein's results showed "overall, they rated their own chances to be above average for positive events and below average for negative events" (Weinstein, 1980, pg. 806).

When discussing "Safety is an abstract concept" West considers:

> "Often the pro-security choice has no visible outcome and there is no visible threat. The reward for being more secure is that nothing bad happens. Safety in this situation is an abstract concept." (West, 2008, pg. 37)

Likewise with "feedback and learning from security-related decisions" West makes the straightforward points that when learning, positive reinforcement is used when something right is done, and negative reinforcement is used when something is done incorrectly. West then points out however that the positive reinforcement in computer security is "nothing bad happens" and the negative reinforcement is "something bad might happen at some stage in the future". The lack of clarity as to what will happen and the lack of immediacy mean users would be less likely to be concerned by the risks of data remanence.

2.5.1 Relevance to Data Remanence

There are clearly a number of psychological reasons why data remanence is still an issue.

Optimism bias with data remanence means that the seller either does not know or does not believe that other people buy second hand devices for the purpose of finding identifying information or commercially sensitive information. In the case where the seller is aware that this is a possibility, the optimism bias suggests the seller believes it is more likely someone purchasing a second hand device for the purpose of finding information would buy someone else's device and not theirs.

Prospect theory in the case of data remanence means sellers gamble that a very bad thing might happen to them, or nothing will happen rather than take the guaranteed loss of time to sanitise their hard drive. This theory is even more applicable when applying the "safety as an abstract concept" argument that people will not know their data have been compromised until something bad actually happens. With data remanence the original owners are unlikely to ever know if someone else has examined the former hard drive for remnant data.

2.6 Risks arising from Data Remanence

The actual remaining data on second hand hard drives can be as harmless as some holiday snapshots of a husband and wife, or as damaging as holiday snaps of a husband and his

mistress. This section will demonstrate some of the types of risk that selling a second hand drive can have. Depending on the jurisdiction and role of the hard drive seller there may also be legal requirements for the seller to properly remove all data.

2.6.1 Legal Requirements: New Zealand Privacy Act

Hooper and Evans (2010) present a guide to the New Zealand Privacy Act of 1993 from the perspective of social networking services. However as those authors note the principles of the Act are written in a "technology neutral way" so they can apply to both Internet and real world situations. (Hooper & Evans, 2010, pg. 123)

The New Zealand Privacy Act outlines what data may be collected about individuals, how data may be collected, how data should be stored and disposed of, how individuals may check data held on them and correct it, and mechanisms to deal with issues arising from breaches of the Act. The Privacy Commission is the first organisation that New Zealanders should contact if they believe their privacy has been breached. The privacy commissioner will then investigate and make a non-binding recommendation. The commissioner does not award damages or inflict fines but can advise if the individual should proceed to the Human Rights commissioner who does have the ability to award damages.

The New Zealand Privacy Commission (2011) outlines the 12 principles of Privacy as defined in the Privacy Act and put them into easier to understand language. Principle 5 (and to some degree Principle 11) relate directly to this research.

Principle 5 is (New Zealand Privacy Commission, 2011a):

> "An agency that holds personal information shall ensure -
>
> (a) that the information is protected, by such security safeguards as it is reasonable in the circumstances to take, against -
>
> (i) loss; and
>
> (ii) access, use, modification, or disclosure, except with the authority of the agency that holds the information; and
>
> (iii) other misuse; and
>
> (b) that if it is necessary for the information to be given to a person in connection with the provision of a service to the agency, everything reasonably within the power of the agency is done to prevent unauthorised use or unauthorised disclosure of the information."

Principle 11 outlines situations where agencies may disclose information and includes (New Zealand Privacy Commission, 2011b):

> (b) that the source of the information is a publicly available publication; or
>
> (c) that the disclosure is to the individual concerned; or
>
> (d) that the disclosure is authorised by the individual concerned; or
>
> (f) that the disclosure of the information is necessary to prevent or lessen a serious and imminent threat to -
>
> (i) public health or public safety; or
>
> (ii) the life or health of the individual concerned or another individual;

Principle 5 is clear on requiring that the agency prevents unauthorised use or unauthorised disclosure, it does assume that properly wiping data from storage devices is reasonably within the power of the agency possessing the storage device. The Government Communications Security Bureau Manual (New Zealand Government Communications Security Bureau, 2010) does provide guidelines for properly wiping data but that manual tends to be aimed at government agencies rather than private enterprise. Conversely there are a number of applications available that companies can use such as scrub3.exe (see Figure 7 on page 83) that can easily wipe the storage device for them.

2.6.2 Espionage

Jones (2008) considers the different contexts for espionage. Jones defines military or political espionage as "one country spying on another" (Jones, 2008, pg. 4) while the industrial form of espionage is defined as "spying for commercial purposes" and may be conducted by "governments, companies and by other types of private organisations such as pressure groups" (Jones, 2008, pg. 7). With regards to industrial espionage Jones describes the main goal as using stealing or copying trade secrets or confidential/valuable information (Jones, 2008, pg. 7).

One example Jones gives is of three people who were convicted for stealing trade secrets from Coca-Cola and trying to sell them to Pepsi. Pepsi however notified the American Federal Bureau of Investigation (FBI) who in turn set up a sting operation to catch the three people (Jones, 2008, pg. 7).

Jones (2005) also discusses industrial espionage from the perspective of some of the remnant data found. In particular "directories of staff, staff profiles and business plans, all of which would be of high potential value to anyone outside the organization." (Jones, 2005, pg. 7)

One important thing to consider is that in 2005 both social media and the ability to search the Internet were not nearly as effective and efficient as they are now, so obtaining staff profiles and directories may not necessarily have been as easy then as it is now. Business plans and other financially oriented documents that could provide competitive advantage however are still typically difficult to find online.

While there do not appear to be any published accounts of data remanence being used for industrial espionage, it is possible that data remanence is a cause of"unknown/unexplainable" espionage, which is where a specific person or cause cannot be tied to the act committed. This also makes sense when considering Garfinkel and Shelat's (2003) hypothesis that people are looking for data but not publicising it. In other words, a spy is not likely to reveal that a company is not properly wiping their hard drives before sale as this would take away the spy's source for obtaining the information.

2.6.3 Identify Theft

Perl (2003) considers there to be three types of Identity theft: "financial identity theft", "non-financial identity theft" and "criminal record identity theft".

2.6.3.1 Financial Identity Theft

Perl describes financial identity theft as financially motivated. It includes using information to directly withdraw money from the victim's bank account, to obtaining new services such as a bank account, line of credit, or a credit card (typically the identity thief will default on these), and in some cases filing for bankruptcy in the victim's name (Perl, 2003, pg. 177). While there is the direct problem of the victim losing money or having their credit rating ruined there is also the secondary problem that the consequences may not be detected for a long period of time (credit card statements often are only sent monthly) which can make it difficult to determine who committed the crime or the method the criminal used to acquire the identity theft information (Perl, 2003, pg. 178).

Anderson et al. (2008) argue that financial identity theft is possible because sellers sell based on a "promise to pay", which assumes that the buyer can pay by providing information about a specific account or a credit history. The identity theft therefore happens when the thief is able to provide enough correct information so that they can "acquire goods while attributing the charge to another person's account."(Anderson et al., 2008, pg. 171).

Results from the Bureau of Justice Statistics (BJS) reported by Langston (2011) who states "In 2010, 7.0% of households in the United States, or about 8.6 million households, had at least one member age 12 or older who experienced one or more types of identity theft victimization" (Langston, 2011, pg. 1). Further results from that report show that general identity theft has been reported more frequently since 2005 (from approximately 5.5% to the 7.0% reported in 2010). For the purposes of their study, BJS defined identity theft as "the unauthorized use or attempted misuse of an existing credit card or other existing account, the misuse of personal information to open a new account or for another fraudulent purpose, or a combination of these types of misuse" (Langston, 2011, pg. 1). This means that they were focusing on financial identity theft above other forms.

However it is not clear whether New Zealand is likely to have the same risks from identity theft as the United States. Johnson (2009) examines identity fraud in New Zealand.
"To date, in New Zealand, there appears to have been no specific research conducted in the identity fraud field, aside from the development of the Department of Internal Affairs led *Evidence of Identity Standard*" (Johnson, 2009, pg. 39). The Department of Internal Affairs (2011a) *Evidence of Identity Standard* is a part of their *Authentication Suite* and is described further by Gray (2010). Gray covers many topics in identity fraud as it relates to New Zealand and ways to mitigate the associated risks. The *Evidence of Identity Standard* itself outlines three main requirements as part of the evidence of identity (New Zealand Department of Internal Affairs, 2011a):

> **"1. Evidence that the claimed identity is valid** – i.e. that the person was born and, if so, that the owner of the identity is still alive.
>
> **2. Evidence that the presenter links to the claimed identity** – i.e. that the person claiming the identity is who they say they are and that they are the only claimant of the Identity.
>
> **3. Evidence that the presenter uses the claimed identity** – i.e. that the claimant is operating under this identity in the community."

Hooper and Evans (2010, pg. 126) state:

> "In New Zealand, the lack of a single unique identifier, such as a social security number or national identity number, means that New Zealand citizens are less vulnerable to privacy invasions and identity theft than their counterparts in the United States, for example."

Attempting to find adequate and accurate figures for identity theft in New Zealand proved difficult. One estimate by the Department of Internal Affairs (New Zealand Department of Internal Affairs, 2011b) calculates the identity theft rate as 3.3% of the total population of approximately 4 million. Assuming that if that estimate is only for financial identity theft data then the New Zealand rate of identity theft is approximately half the value given for America in the BJS 2010 statistics (Langston, 2011).

2.6.3.2 Non Financial Identity Theft

The next form of identity theft considered by Perl is "non financial identity theft". Perl states that other motivations for identity theft that do not have a direct or instant financial payoff include obtaining government documents such as passports to gain unlawful entry into a country, which then could be used for terrorist attacks, or acts of espionage (Perl, 2003, pp. 178-179).

2.6.3.3 Criminal Record Identity Theft

The third form of identity theft considered by Perl is "criminal record identity theft". This form is where a criminal uses a stolen identity to cover up other crimes. As such Perl states that this form of identity theft is the "the worst-case scenario of identity theft" as it "allows criminals to stay on the street without being arrested", and because of the challenges associated with the victim restoring their identity, proving that they had not committed the crimes the thief committed in their name and then correcting their records to show they did not commit them (Perl, 2003, pp. 180-181).

2.6.3.4 Potential Identity Theft and Data Remanence

While there is little direct evidence of data remanence being the root cause of specific instances of identity theft, there are many cases cited where the root cause cannot be determined. The following are a number of examples presented by various researchers

regarding data remanence and how the information they found either partially or fully matches the previous descriptions of identity theft.

Some examples from the 2009 consortium report (Jones et al., 2010, pg. 101) include:

- "It disclosed the personal phone numbers of members of parliament
- Australian accounting firm. It contained tax file numbers, customer data, letters of advice, letters of tax rulings, peoples personal income, personal asset base and company and business data
- significant quantity of patient data including tests, test outcomes, letters of advice and of course extensive personal medical histories of ongoing patients.
- On one disk the documents recovered revealed the user's name, occupation, the specific organisation and address at which they worked, as well as telephone numbers, email addresses, a complete employment history and details of their personal education. A large amount of personal email was also recovered."

Medlin and Cazier (2011, pg. 32) reported that:

> "Another hard drive contained information related to tax returns which contained individual client's names, addresses, phone numbers, social security numbers, and date of birth, almost everything one would need to assume another person's identity."

El Emam et al. (2007, pg. 7) were primarily concerned with finding personal health information but reported that some of the identifying information they found included:

- "personal budgets, salary information, tax returns and completed tax filing forms
- payroll records of employees, including addresses, dates of birth, and social insurance numbers
- police record checks"

Kwon et al. (2006, pg. 81) provided a matrix of low risk to high risk identifying information that may have been found and then elaborate with specific examples they found such as (Kwon, et al., 2006, pg. 85):

- "In another three disks used at certain insurance company in Korea, we detected customers' names and resident registration numbers of approximately 2,650 persons

- One of disks contained membership applicant information for an amateur anglers' club, which included resident registration number, address, and so forth for each member"

2.6.4 Blackmail

As seen in the preceding examples, information gained from data remanence could be used directly for financial gain. Blackmail however, uses the threat of not revealing this information to other parties in exchange for payment. One possibility is finding objectionable or illegal material (such as child pornography), tracing it back to the former owner and threatening to reveal it to the police or people that the former owner knows. Other possibilities may include finding legal erotica the owner may have created (pictures or sex tapes) and threatening to post it publicly. The challenges for the blackmailer would be accurately identifying the former owner, and determining whether they can and will pay the blackmail amount.

2.6.5 Stalking

Stalking can be defined in a number of ways but typically it involves "willful[sic] and repeated following, watching and/or harassing of another person" (Wikipedia, 2012). Medlin and Cazier state that (2011, pg. 28):

> "It should also be noted that it is not just the data that contains personal information that can be exploited. Files containing video and audio footage, blogs, diaries, and instant messenger conversations can prove to be equally damaging and more easily exploited, especially if it can be linked to an individual. Calendars as well as address books provide routines and places that may be used to stalk an individual."

2.7 Summary

As the background material into data storage devices and data remanence shows there are a number of considerations to what the risks are to people from not properly wiping their devices before disposing of them, why the data are still there (both technically and psychologically) and what researchers and investigators need to consider before conducting data remanence research. The next chapter considers the literature of specific data remanence research in order to build a methodology as described in Chapter 4 for appropriate data remanence research in New Zealand.

Chapter 3: Literature Review

3.1 Introduction

This chapter is firstly divided into data remanence for hard drives and then other data storage devices. The hard drive research is further broken down into sections covering the pilot studies such as Garfinkel and Shelat (2003), Valli (2004) and Jones (2005), then the work of the consortium who compare results from a number of countries and focus mainly on "how many drives are wiped?" and "how many companies and/or individuals be identified from any data remaining?". Then the section on hard drives considers the researchers who had specialised research questions such as El Emam et al. (Primary Health Information), Kwon et al. and Medlin and Cazier who had a focus on identity theft.

Section 3.5 considers data remanence for non hard drive data sources, then auxiliary research such as Sutherland and Mee (2006) is briefly considered, and the chapter concludes with a summary of the sorts of research questions asked and leads into the methodology for this thesis in the next chapter.

3.2 Pilot studies

The main pilot studies into hard drive data remanence focused on different research questions and the published results had different reporting styles and presented different levels of explanation of their methodologies.

One of the first major works in the quantitative study of computer data remanence was Garkinkel and Shelat (2003).

They provide a key example of why data remanence is an important issue. The United States Veterans Administration Medical Centre in Indianapolis retired 139 computers and because the drives had not been sanitised medical records and credit card details were found (Garfinkel & Shelat, 2003). They also relate five more similar cases where data considered to be sensitive had been recovered because the drives had not been sanitised. Garfinkel and Shelat noted that the cases had strong similarities but in their opinion the cases were also scarce.

They then formulated three hypotheses (Garfinkel & Shelat, 2003, pg. 18):

1. Disclosures of this type are exceedingly rare
2. Confidential information is disclosed so often on retired systems that such events are simply not newsworthy
3. Used equipment is awash with confidential information, but nobody is looking for it—or else there are people looking, but they are not publicizing the fact.

To test their hypotheses Garfinkel and Shelat purchased 158 hard drives from November 2000 to August 2002 purchasing from both online sites such as eBay and from computer stores that sell used hardware (Garfinkel & Shelat, 2003, pg. 24).

The strengths of Garfinkel and Shelat's work include the comprehensive nature of their own prior work. They created a Sanitisation taxonomy, list various tools (most of which are still in use in 2011) and suggest reasons why users do not sanitise their disks before selling them. The methodology is reasonably straightforward and has been used by others in the field of data remanence, including this author, with some modifications.

Some of the challenges arising from their work however include not having tables outlining exactly the number of drives examined, how many were unreadable, how many were wiped, and how many had identifying information. This can make it difficult to understand the depth of the problem and also makes it difficult to compare results between different researchers. One of the tables that Garfinkel and Shelat do include is a list of Microsoft file types such as PowerPoint, and Word and the number of hard drives that have those files on them, but it is unclear what the total number of drives is with those files (for example a drive may contain both PowerPoint and Word files) and the nature of those files. Garfinkel and Shelat also mention finding pornography (Garfinkel & Shelat, 2003, pg. 25) on the hard drives, but they do not mention how many hard drives had pornography on them, if the pornography would be classified as legal erotica or illicit/illegal, and if there were any legal issues arising from that material. This in part is because their methodology lacks a statement as to what would happen if they did find illegal material. Presumably it would be handed over to the authorities, but this is not explicitly stated.

Valli (2004, pg. 124) initially researched eleven computers that had been purchased from six different auctions in Australia. Valli used a five step methodology for determining the contents of the hard drive. His five step methodology has been included in Appendix E. If

nothing could be found after the five steps, the drive was considered to be "secure". Only one of the eleven drives had been securely erased (Valli, 2004, pp. 126-127).

He concluded that nine of the drives had been formatted and were most likely from corporate environments based on the data found. Valli also discusses various aspects of the data recovered from the hard drives by providing examples, and the possible profiles of former owners that could be determined from the data.

One of the strengths of Valli's initial research was to provide a clear step by step methodology, as well as describe the specific tools used to conduct the research (*Foremost* and *Autopsy*). Given the nature of the data found, it was also assumedly easy for Valli to classify the data as "individually identifying" and "company identifying" (though not referred to as such in Valli's work at that time).

While Valli makes it clear that it is a pilot study (hence 11 drives examined), comments on costs, time required or recommended sample sizes for further work could have been useful for other researchers. The research did not state the type of interfaces used on the hard drives, so there is no distinction between "IDE", "SATA" and others. This becomes relevant when compared with later research by Valli and Woodward (2007) who used a "targeted" approach of purchasing SCSI drives and laptop drives with the goal of finding corporate and/or government drives. (Valli & Woodward, 2007, pg. 219). The targeted approach is consistent with Jones' (Jones, 2009) assertion that companies tend to use SCSI drives.

Jones (2005) conducted research in a similar fashion to both Garfinkel and Shelat, and Valli (2004) for the United Kingdom. Rather than using tables, Jones presents the data in small sidebar paragraphs which makes it easy to find and informative as he describes each finding. Examples include (Jones, 2005, pp. 5-6):

"**Identifiable to a user:** (49/92) 53 percent of the disks contained identifiable usernames. A number of these disks contained multiple usernames."
Financial information: (18/92) 20 percent of the disks contained financial information relating to the organizations, including staff salary details, sales receipts and profit and loss reports.
"**Network information:** (7/92) 8 percent of the disks contained details regarding the network infrastructure of the organizations. This information included server names, proxy numbers and other IP number details."

Overall the reporting style is interesting and informative but can take some extra work to determine the main values, as opposed to a table approach. Jones also includes specific details about certain drives found.

As a summary of the pilot studies there were two rather different types of question asked by the researchers, firstly Garfinkel and Shelat were more interested in the types of data that could be found, while Valli, and Jones were more interested in whether the drives were wiped or who could be identified from the information found.

3.3 Longer term and comparable research

After the joint publication of the comparison of the 2005 second hand hard drive data remanence results by Jones and Valli (Valli & Jones, 2005), a research consortium was formed, that initially included universities from Australia, the United Kingdom, the United States of America and Germany. Hard drives from France were included in the analysis from 2008 and in the 2009 research it is mentioned that Khalifa University of Science Technology and Research contributed to the analysis of the disks (Jones, et al., 2010, pg. 10).

This suggests that drives may have been purchased online from various countries, marked with the source country and then shipped to a third party for analysis or that much like Garfinkel (2010), the drives were imaged in their home country and those images were then analysed in another country. In the former case, it is possible that the extra distance the hard drives are sent may result in the drives being unreadable and in both cases the issue of what is illicit or illegal in one jurisdiction may not be illegal in another jurisdiction.

The reporting and methodology were similar to the previous research by Valli (2004) with prominent and relevant hard drives being mentioned in more detail. Each year's publication also discussed aspects of why some values may have differed between countries such as the difference between unreadable drives in 2006, with Germany having 72% unreadable drives vs. Australia only having 6% unreadable drives (Jones et al., 2006, pg. 28). Appendix I presents the four years of consortium results from 2006 to 2009.

The presentation of their reports changed slightly in 2008 with the additional classification and reporting of drives that were "formatted" which includes "removing data by deletion, formatting or reinstalling an operating system" (Jones, Dardick, et al., 2009, pg. 165). For the purposes of this thesis, "reformatted" is the term to describe such drives. It is possible for a

hard drive to have both individual identifying information and company identifying information on it. Additionally if the tools used for the forensic analysis allow for data carving then a drive could be classified as formatted and identifying information could also be found. The totals for each year are summarised in Table 1 (based on Tables 28-31 in Appendix I) to enable comparisons between years and also between this research and previous research. One key observation is that the percentages do not necessarily add up to 100%. The consortium count the unreadable drives, and then remove that count when calculating the percentages for wiped drives. They then calculate percentages based on the "remaining" total for identifying information. As this can be confusing, this thesis uses a slightly different reporting style which is addressed in Section 4.10 on page 96.

Table 1 Amalgamation of consortium 2006-2009 results

Year	Total	Unreadable	Wiped	Remaining	Company Identifying	Individual Identifying	Formatted *	Illicit *
2006	317	126 40%	61 32%	130	46 35%	41 32%		17 13%
2007	300	112 37%	62 33%	126	51 40%	56 44%		22 17%
2008	338	119 35%	61 28%	158	62 39%	55 35%	83 53%	19 12%
2009	346	94 27%	89 35%	163	65 40%	64 39%	34 21%	8 5%

* These values were not consistently reported for each year or country and the percentages may be higher in reality.

Overall the consortium has provided the most consistent investigation into data remanence for hard drives, with over 300 hard drives analysed each year. Unfortunately there is no

specific analysis given as to why the "unreadable" rate has decreased over time. This may be due to the better technology in newer hard drives making them more reliable. The other observations such as wiped, and identifying show less clear trends.

3.4 Specialised research questions

Not all researchers into data remanence are strictly interested in the number of drives in a sample that are wiped, or if companies or individuals could be identified. Researchers such as El Emam et al. (2007) and Medlin and Cazier (2011) have asked different more specialised questions. Kwon et al. (2006) have been included as they have taken a mixed approach to the previous research.

3.4.1 Primary Health Information

El Emam, et al. (2007) focused on the personal health information (PHI) on second hand hard drives sourced in Canada and conducted their research in a generally similar way to the consortium.

The research however differed in that the researchers make very clear the selection method used for sourcing their drives (by listing all Canadian second hand hard drive vendors and randomly selecting a sample of those vendors) and the statistical methods used to determine that they needed to select at least 60 hard drives (El Emam, et al., 2007, pg. 26). Additionally those researchers were only concerned with "functional" hard drives and "all non-functional drives were returned and replaced." (El Emam, et al., 2007, pg. 25). For comparative reasons it would have been of interest if they had reported how many unreadable drives were discarded as per their methodology. Finally they also excluded drives for a number of other reasons, which means their research could also be classified as "targeted" (similar to Valli's and Woodward's (2007) approach of selecting SCSI drives to target corporate data).

For example El Emam, et al. excluded large disk drives as they felt that these might be from servers that contained large databases of personal information. Their methodology does not state the interface(s) of the drives they examined however so it is unclear if they only examined IDE drives or not. They also excluded drives that vendors did not wish to sell to them, typically due to the geographic distances involved. This led to their observation that "We suspect that vendors with drives containing un-wiped PI and/or PHI were less likely to sell us the equipment." (El Emam, et al., 2007, pg. 35).

El Emam et al. also had three special protocols as part of their research which primarily dealt with the legal issues surrounding data remanence. First they did not examine any image files (to avoid inappropriate or obscene material), and had a member of the research team screen the hard drives for suggestive file names and flag drives that contained such names for special consideration. Second, they made it explicitly clear that any illegal materials found would result in the police being contacted, and the third protocol was if "particularly sensitive" personal information or personal health information was found, the appropriate privacy commissioner would be contacted (El Emam, et al., 2007, pg. 28).

Additionally they provide clear definitions of personally identifying information and how they searched for it. As part of the searching methodology one researcher examined all 60 hard drives and then a second researcher examined a subset of the drives to ensure that the drives had been consistently classified (El Emam, et al., 2007, pg. 27).

They went a step further once they had analysed the data, which was to try to match the geographical location of the "individual identifying" information they found back to the owner and to the vendor who sold the drive second hand. While they were reasonably successful in that (26/39 drives) (El Emam, et al., 2007, pg. 29) and it was relevant to their research as it related specifically to health information and potentially medical practitioners, the author considered it outside the scope of this research. It was interesting to note that the owners for 22 of these 26 drives were in the same province as the vendor. The other four drives were US-owned but were sold by Ontario vendors (El Emam, et al., 2007, pg. 29).

This suggests that there could be some interesting follow-up work for the consortium members such as France and Germany in determining whether any of their hard drives came from bordering countries or if the drives were all internally sourced. As an island country this may be less of an issue for New Zealand.

EL Emam et al. compared the state of the drive as described by the vendors such as "All data wiped" with what they actually found. The methodology for this thesis also considered that issue and determined it would make more sense and better use of resources to purchase drives where no explicit statement of the drive contents had been made other than the drive was second hand or used.

They also differentiated between "blank" and "scrubbed to a DoD*20.22-M Standard", and also "repartitioned" and "formatted" in reporting their results, and due to rejecting outright

unreadable drives it means the Canadian results are incompatible with the consortium reporting style. El Emam et al. do however report that overall they found 28/60 (47% of the total) drives with individually identifying information, 5/60 (8% of the total) drives had PHI about the person they believed to be the owner of the hard drive and 6/60 (10% of the total) drives had PHI of people other than the owner of the hard drive (El Emam, et al., 2007, pg. 30). They additionally state that "Our results indicate that not as much health information is leaking as other types of information, such as financial and legal information" (El Emam, et al., 2007, pg. 31)

The strengths of El Emam et al.'s research are many when considering the design of data remanence research. The recommendation to pre-screen drives and also to have a second researcher examine a subset of drives to ensure the drives have been correctly categorised adds to the repeatability of the research and validity of the results. Including clear protocols for how to deal with illegal material is also something not included in several of the other methodologies encountered. Defining an appropriate number of readable drives to examine also adds a dimension to their research which other researchers do not explicitly state.

If El Emam et al. had included the number of unreadable drives in their sample it would have made the results comparable with the consortium results. Overall El Emam et al.'s research provides many features that make sense to include in data remanence research.

3.4.2 Identity Theft

Medlin and Cazier (2011) focused on identify theft on second hand hard drives, but their experimental methodology used live forensics and did not use sophisticated forensics tools. This allowed them to examine the drives as if they were typical users.

Medlin and Cazier's methodology was to purchase hard drives from thrift stores (aka second hand stores) and also hard drives that had been donated by students in America in 2007. They collected a total of 55 drives. The methodology used for finding information involved a variant of live forensics as "only those hard drives that either booted, on their own or with a Live CD, were examined in depth." (Medlin & Cazier, 2011, pg. 31). From a chain of custody perspective (see Section 2.4.3 on page 26) this would typically raise concerns if illegal material was found and is therefore inconsistent with previous methodologies of research that Medlin and Cazier had referenced.

Of the 55 drives examined by Medlin and Cazier, 11 were classified as unreadable (Medlin & Cazier, 2011, pg. 31). Because of their methodology it is unclear if their definition of "unreadable" includes drives that were wiped or reformatted (as the hard drive would appear to have nothing to a typical user), or hard drives that would be classified as unreadable due to being broken, not starting up or not being detected by the computer the hard drive had been put into.

They report the number of instances of specific types of information useful for identify theft such as full names, phone numbers, addresses, other financial documents, social security numbers, bank accounts, tax returns or related tax information, credit card information, debit card information, PIN numbers, and wills (Medlin & Cazier, 2011, pg. 31). The inclusion of "wills" may seem odd; they reported finding 3% in that category with an explanation that it occurred on only one of the hard drives. There is also no definitive statement on how many of the drives contained identifying information and of those, how many contained information useful for identify theft. They do include counts of types of files that were used to find the information such as "cookies", "Temporary internet files", "pictures" and the like (Medlin & Cazier, 2011, pg. 32).

One of the strengths of the research by Medlin and Cazier therefore is the definitions of identity theft and ways that it can be found by typical users when considering second hand hard drives that work by being plugged into a desktop. This clearly outlines one of the biggest threats of data remanence and also uses a low tech approach to emulating typical users. Of concern, much like Garfinkel and Shelat (2004), is that the issue of what would happen if pornographic or illegal material was found was not addressed, which is reinforced by the methodology not appearing to consider the chain of custody.

3.4.3 A mixed approach

Kwon et al. (2006) considered data privacy issues on second hand hard drives using a methodology similar to the consortium's but using a reporting style more consistent with Garfinkel and Shelat. Kwon et al. (2006, pg. 82) purchased a total of 55 hard drives of which 17 were found to be unreadable. Kwon et al.'s research differed slightly from the consortium as it also considered a count of specific pieces of identifying information found (similar to the counts presented by Garfinkel and Shelat, Medlin and Cazier but different in context and type of count presented).

Kwon et al. state they found identifying information related to over 4000 individuals, however their results include one hard drive from an insurance company which accounted for almost 3000 of those individuals (Kwon, et al., 2006, pp. 83-84).

That one drive represents approximately 65% of all the identifying individuals found. Kwon et al. also used a metric of "resident registration number" which suggests that merely finding someone's name might not be classified as "identifying information" for their research and the count of matched names and matched resident registration numbers might be a better metric. As part of their prior work section they describe both Garfinkel and Shelat's work as well as Jones's 2006 work, and include a table comparing those results (Kwon, et al., 2006, pg. 78).

Kwon et al.'s results were then published in two forms. The first form was a table similar to the one used to compare Garfinkel and Shelat's work. The second form was two tables that outline specific cases along with the above mentioned counts of "Num. Of Exposed People", and notes on the source of drive and other identifying information.

A strength of Kwon et al. is the framework of describing information as it relates to identity and therefore the potential threat from identity theft and other aspects of data remanence. Much like Medlin and Cazier, Kwon et al. did not appear to make a bitwise copy of their hard drives before analyzing them nor did they outline what would happen if illicit or illegal material was found. A secondary issue is the lack of consideration to that one hard drive from the insurance company contributed 65% of the results where in other research it would likely be considered an outlier and addressed as such.

3.5 Data remanence for other data storage devices

There appeared to be less data remanence research regarding data sources that are not hard drives. As shown in Section 2.2.4 (on page 17) one of the main reasons for this is that while hard drives typically require few cables (due to the standardisation of interfaces) smart phones can often require many different types of cables. A second challenge is that while many hard drive data remanence researchers can use dead forensics, typically mobile/smart phone researchers may have to use live forensics as explained in Section 2.4.2.2 on page 26.

3.5.1 Mobile/Smart phones

Glisson et al. (2011) published results on data remanence for mobile phones, using a similar methodology as hard drive data remanence researchers, by purchasing second hand mobile

phones and then analysing those phones. Glisson et al. (2011, pg. 338) claim that they "provide the first empirical report on the extent of artifacts retained on mobile devices when they are passed into secondary markets"

They noted a number of limitations on their research such as they only examined phones from the United Kingdom, they could only purchase phones that were compatible with the forensic toolkits they had available and that they only investigated 49 phones of the approximately 80 million phone subscriptions in the UK (Glisson, et al., 2011, pg. 339). As with other consortium publications, sample size considerations and margins of error were not reported. An interesting element of their reporting style was they also included a banded breakdown of the prices of the cell phones with the majority costing less than 60 pounds sterling each. This is interesting for future research as it can potentially assist with budgeting and also by highlighting that owners of more expensive or less expensive brands of mobile may be more likely to have data remanence issues. That would be a fascinating research question that is not as easy to study for hard drives, as the entire system the user had would need to be purchased, and the price analysed to when the system was originally purchased.

One of the interesting outcomes of Glisson et al.'s research is that when they considered the "how many drives are wiped?" equivalent for their research, that none of the phones had been wiped before being sold.

3.5.2 USB storage devices

Jones, Valli et al. (2009) have conducted research into data remanence in USB storage devices. That research followed the same general methodology and aims as the hard drive data remanence research with the consortium. They defined USB storage devices as being "a thumb drive, a keychain drive and a flash drive" (Jones, Valli, et al., 2009, pg. 1) making it clear that it did not include USB external hard drives which are typically much larger capacity and physical size.

They examined 43 USB storage devices and observed that although the USB devices had less storage capacity for data, in general the USB devices had much more identifying data than their hard drive counterparts based on results from other consortium research. Their research is unclear what the capacities (in megabytes or gigabytes) of the USB storage devices were (which is addressed by Chaerani et al. (2011)) and where the devices were sourced from.

Additionally it is unclear if the devices were sourced individually or whether there are vendors that sell second hand USB storage devices in bulk.

Chaerani et al. (2011) conducted USB storage device research in 2011. Their methodology was similar to Jones, Valli et al. (2009) but the key differences include the reporting and presentation of further analysis. They included a breakdown of the counts and capacities of the storage devices (ranging from 128 megabyte to 4 gigabyte) and showed the total counts of the files found similar to Medlin and Cazier (2011). An addition to the reporting style was that they included a table of counts where they believed specific crimes could be committed such as "identity theft", "fraud", "blackmail" and "robbery". The latter was identified because the former owner of a thumb drive had included specific vacation plans, the timeframe for that vacation and other details about where their house was located which a criminal could potentially use to know when the former owners were going to be away for an extended period of time.

Chaerani et al.'s research only studied 20 devices. This is acknowledged in their conclusions, and it is noted that further work in the field of USB storage device data remanence is needed. As Chaerani et al.'s paper was published in December 2011 it was considered too late to code the results from this research using the "possible crimes" approach that they used, however it is definitely an aspect under consideration for future research projects.

3.6 Techniques to automate Data Remanence research

Automating forensic analysis is one way to improve data remanence research. Fragkos et al., (2006) considered that based on what they describe as the "forensic race" of the drive (predominantly the drive capacity and apparent data on the drive) and other factors it may be possible to calculate the time it would take to analyse the drive and a schedule could be built around that. In other words, a very large drive with little to no data should take less time than a small capacity drive that is full of data. Their methodology has three steps which are excluding the faulty discs, automating a process to examine the wiped drives and then using potentially four different tools to examine the remaining drives (Fragkos et al., 2006). They also suggest that there can be benefits to grouping drives from the same source if possible, and by examining based on the amount of content rather than the total capacity (largest to smallest). Fragkos et al. in their conclusions section stated that "We believe that in the near

future a system could be built that will be able to perform all the steps presented here automatically" (Fragkos et al., 2006, pg. 103).

Garfinkel (2009) considered the automation of forensic processing and noted a lack of research in that area despite a perceived need for automated tools. He then acknowledged that *Encase* and *FTK* are two of the most commonly used forensic analysis tools, and while *Encase* does have a scripting language, it is proprietary to *Encase*, and often requires expensive training to learn a specific skill that is not easily transferable. Garfinkel also notes that *Pyflag* is another forensics tool with scripting capacity but it also suffers from being difficult to deploy and overly complex (Garfinkel, 2009). Garfinkel's goal was to develop tools for the automation of disk forensic processing. He presents three applications the first being a tool to generate data and metadata about the filesystem (and inodes/files) on a disk in order to map the drive, the second being a tool that can redact parts of (disk) images and the third being a stand-alone kiosk to safely transfer data from possibly virus infected USB drives.

One of the advantages of the first application of Garfinkel's (2009) approach is the remote access and analysis that is possible for data remanence. As Garfinkel notes, there were two terabytes of data available for their research and while writing all of the data to four large 500 GB capacity hard drives was a possibility, there was the practical alternative of setting up an SSL secured website and having the researchers in other locations log in with passwords and use the automated tools over the Internet. The XML and tools meant that only a small subset of the terabytes of data needed to be examined.

The second application is clearly useful as it means potentially sensitive information such as a user's real name can be overwritten or redacted with "John Doe" or similar, and the third application also could be useful for data remanence research as it provides a dead forensics approach to reading data from a USB thumb drive.

From the perspective of this research however, much like *Pyflag*, the time required to learn and deploy Garfinkel's tool and determine its effectiveness was considered to be too long, as *Encase* and *FTK* were already available to the author. Additionally as this research was conducted only on-site at the University of Otago, the collaborative nature of using the XML for Internet access was not required.

3.7 Auxiliary research

There has been some research that would be considered auxiliary to data remanence, including Thomas and Tryfonas (2007) which summarises other experiments, and other research which considers the challenges of data remanence from other perspectives. Sutherland and Mee (2006) used a survey (rather than physically inspecting hardware from schools) to measure how schools dispose of their hardware. Sutherland and Mee note they had a small sample size based on 24 returned questionnaires (Sutherland & Mee, 2006, pg. 229) but were able to observe that 10 of the 24 responding schools were not aware of procedures to properly remove the data before disposing of their equipment. It was also observed that a number of schools formatted and re-installed operating systems as a mechanism to remove data which is generally not enough to remove the risks from data remanence. Johnson et al. (2009) surveyed schools in New Zealand and merely asked how schools disposed of their hardware (donations, landfill, recycling centres) without asking what schools did before disposing of that equipment.

3.8 Literature review summary

This chapter has presented the literature review of data remanence experiments.

Researchers have typically asked "how many drives are wiped?" (or a variant thereof) to find out how many drives in a sample have been properly prepared for disposal. Valli, Jones and the consortium have also asked "How many drives had information that could identify companies and/or individuals from the remnant data found?" as this can suggest risks to those who do not prepare their drives properly. El Emam et al. take this a step further by investigating if primary health information could be found, and Medlin and Cazier examined the data found for identity theft risks. Chaerani et al. examined data found on USB storage devices in a similar way to Medlin and Cazier but included categories for other potential crimes such as blackmail and robbery.

Auxiliary research such as Sutherland and Mee (2006) raised the potential question of "Which companies have more issues with data remanence than others?" and that has not been directly answered by the consortium or other researchers. That may be because other researchers were not able to get enough data for suitable sample sizes, or that they had not considered it.

As seen in the non-hard drive research most of the research would be considered as pilot studies and addresses the technical issues that make data remanence for those devices difficult. Glisson et al. (2011) consider their research to "provide the first empirical report on the extent of artifacts retained on mobile devices when they are passed into secondary markets" and studied 49 devices of a potential 80 million devices.

The challenges outlined include traditional forensics software not necessarily being appropriate to use for mobile phone forensics, likely having to use live forensics rather than dead forensics, and despite the large number of devices available that very specific cabling and hardware may be needed whereas with hard drive data research is much more standardised.

Chapter 4: Methodology

4.1 Introduction

The methodology for the experiments conducted for this thesis was created by examining all the relevant methodologies for computer data remanence as outlined in the previous chapter and attempting to determine the best and most practical elements to implement. Much of the research conducted into data storage devices that were not hard drives has only been published since late 2010 (Owen et al., 2010; Chaerani et al., 2011; Glisson et al., 2011; Grispos et al., 2011) after the initial experiments were designed and conducted for this thesis. Garfinkel (2010) reinforces the decision for this thesis research to only consider hard drives as hard drive operating systems, file formats are well documented and interfaces/cabling has been standardised, as compared with the many different cables that are potentially needed to capture smart phones and the various operating system and architecture issues that can surround those devices.

4.2 Special considerations

Depending on the jurisdiction and the type of research or analysis being conducted there may be special considerations that need to be addressed before the formal methodology is designed. For this thesis, ethical approval was required. Due to the potential for finding material that is legally objectionable in New Zealand protocols were put in place in case that material was found. Securing the data was also considered.

4.2.1 Ethical Approval

The University of Otago "requires that any research involving human participants is conducted in accordance with the highest ethical standards." (University of Otago, 2011). The University of Otago has two categories for considering ethical approval with regards to teaching and research, those categories being "A" and "B". This research was submitted under Category "A" as it meets the following criteria as directly quoted from the University of Otago's "Ethical Practices" page (University of Otago, 2011):

> "Personal information - any information about an individual who may be identifiable from the data once it has been recorded in some lasting and usable format, or from any completed research (Note: this does not include information such as names, addresses, telephone numbers, or other contact details needed for a limited time for

practical purposes but which is unlinked from research data and destroyed once the details are no longer needed);"

Approval was granted by the Ethics Committee for Category A research before the research started and can be found in Appendix D.

4.2.2 Special protocols for objectionable material

As the prior work by El Emam et al. (2007) and others have shown there is always the possibility that "objectionable material" may be found on any of the hard drives. In the New Zealand context "Objectionable Material" is a legal term defined in the *Films, Videos and Publications Act of 1993* (New Zealand Films Videos and Publications Classification Act, 1993):

> "For the purposes of this Act, a publication is objectionable if it describes, depicts, expresses, or otherwise deals with matters such as sex, horror, crime, cruelty, or violence in such a manner that the availability of the publication is likely to be injurious to the public good.".

The Act then further clarifies those definitions with what is commonly known as "Child Pornography" being one of the main examples. The definition of a publication includes the more obvious things such as books but also includes (New Zealand Films Videos and Publications Classification Act, 1993):

> "(d) a thing (including, but not limited to, a disc, or an electronic or computer file) on which is recorded or stored information that, by the use of a computer or other electronic device, is capable of being reproduced or shown as 1 or more (or a combination of 1 or more) images, representations, signs, statements, or words"

For this research if at any time anything that is likely to be deemed "Objectionable Material" is found, research on that particular drive is stopped, and the New Zealand Police were to be contacted. The documentation for that hard drive will then note that the drive had "Objectionable Material" and no further processing would continue. The hard drive would be removed from the sample entirely and considered to have not existed for the purposes of this research.

4.2.3 Securing the data

The author considered a protocol for securing the data which was to encrypt the data found. With regards to this research, given the volume of data (100 hard drives and approximately two terabytes of uncompressed data), it was considered that encryption would add extra steps and require extra time that was not merited. The data were secured in a data cabinet (functionally similar in size, shape and design as a normal safe, approximately 1m by 1m by 1m in dimensions) which required a combination to open, and the data cabinet was in a locked office in a building that required key card access outside normal hours.

4.3 Methodology design

As described by Valli (2004), one of the main considerations for data remanence research is to initially determine what can be recovered and how much effort was required to do that (for example, whether data could be seen by just plugging in the hard drive or if specialized tools were needed). Valli subsequently then made "reasoned evaluations of the general profile of the ex-owner of that drive" and noted that "in-depth analysis of the content was beyond the scope of this research" (Valli, 2004).

Unlike Valli's research, the research in this thesis did not include profiling individuals or organisations, however where such evaluations could be made they were noted. With businesses, a factor in determining if they were identifiable was if their name and industry type could be reasonably determined.

4.3.1 Comparison of methodologies from prior researchers

Table 2 (on the next page) presents elements of the methodologies discussed in the previous chapter that were considered relevant and easily comparable.

"Tools named" indicates whether the researchers provided details of the tools used for the capture and analysis of the data.

"Bitwise copy" indicates whether the researchers captured a bitwise copy of the hard drive before analyzing it.

"Illegal material protocol" indicates whether the researchers had a protocol in place if illegal material was found during the analysis.

"Description of what is identifying information" indicates whether the researchers made it clear (and therefore repeatable) what identifying information meant for their research.

"Methodology for finding identifying information" indicates whether it is clear how the researchers examined the drives for identifying information.

"Sample Size considerations" indicates if the researchers made statements about how they determined the sample size and/or made statements that meant future researchers could determine an appropriate sample size. This is addressed further in Section 4.3.2.2 on page 60.

Table 2 Comparison of different features from published methodologies

	Garfinkel 2003 (Garfinkel & Shelat, 2003)	**Valli 2004** (Valli, 2004)	**Jones 2005** (Jones, 2005)	**Korea 2006** (Kwon, et al., 2006)	**Canada 2007** (El Emam, et al., 2007)	**Medlin 2007** (Medlin & Cazier, 2011)	**Consortium 2009** (Jones, et al., 2010)
Tools named	YES	YES	YES	NO	YES	NO	YES
Bitwise copy	YES	YES	YES	NO	NO	NO	YES
Illegal material protocol	NO	NO	NO	NO	YES	NO	YES
Description of what is identifying information	NO	N/A	YES	YES	YES	N/A	NO
Methodology for finding identifying information	NO	N/A	YES	YES	YES	N/A	NO
Sample Size considerations	NO	NO	NO	NO	YES	NO	NO

Where there is a "NO" it is possible that the factor was considered and had merely not been described rather than ignored. "N/A" means that the element was outside the scope of the study in question.

From Table 2 (on previous page), the author constructed a methodology that included the six elements. This would mean research conducted using this methodology is repeatable, especially as some of the other methodologies used were unclear on a number of points, most notably the sixth element "sample size considerations".

4.3.2 Other considerations from previous methodologies

There are a number of factors from other methodologies that either were not mentioned by those authors or are potentially too subjective to compare easily. Reporting, sample size considerations, research questions and identifying information were four factors considered important enough for further discussion.

4.3.2.1 Reporting

Reporting was excluded from the Table 2 feature comparisons because different researchers had different goals which were reflected in the information that they reported. The reporting style for this thesis and how it differs from the consortium in particular is addressed in Section 4.10 on page 96.

4.3.2.2 Sample size considerations

Sample size considerations as stated is something not usually addressed by other researchers (Table 2). This may be because those researchers assumed the sample size could be determined (or assumed) from the number of drives analysed or that in certain cases the research was considered a pilot study and funding or other resources may not have been available to sample a large number of hard drives. One of the related issues from this is attempting to determine the population that the sample was taken from, and this in turn depends if the sample is from "all secondhand hard drives available" or "only secondhand hard drives that do not make a clear statement about their contents" (generally referred to as "unknown state drives" in this thesis).

Two considerations were then made when determining an appropriate sample size for this thesis. The first was "How many drives would it be possible to acquire in a suitable timeframe?" and the second consideration was "What would be an acceptable margin of error for the sample size?"

As for the first consideration, previous consortium sample sizes (Tables 28-31 in Appendix I) were considered and it showed that while most countries examined up to 50 drives each, the United Kingdom consistently studied over 150 drives. The tools available for this research such as *Tableau* and the Spike Effect program also suggested that analysing 50-150 drives would be achievable (barring any supply or transport issues) in the timeframe.

For the second consideration it was determined that this research would make the initial assumption that the overall market (population) for "unknown state drives" on the secondhand market in New Zealand was "large" (over 5000). Consulting sample size tables such as Conroy (2012, pg. 4) (reproduced in part below as Table 3) shows that a sample size of 100 drives would have a margin of error of approximately ± 10% and that a sample size of 384 would be needed to get the margin of error down to ± 5% for a "Large" population. All confidence levels in this thesis are at 95% unless otherwise stated.

Table 3 Sample Size Example Table

Acceptable Margin of Error	**Size of Population**			
	Large	**5000**	**1000**	**200**
± 20%	24	24	23	22
± 10%	96	94	88	65
± 7.5%	171	165	146	92
± 5%	384	357	278	132
± 3%	1067	880	516	169

As a second observation if the population size was lower than "Large" then the margin of error would also be lower (such as at a population of 1000, the margin of error would be closer to ±9.3% for 100 drives). Therefore the sample size of 100 drives with a margin of error of approximately 10% was considered acceptable with the observation that if 150 or more drives were examined or the population was smaller then the margin of error would be lower.

The other issue when dealing with margins of error with this research is that some sample sizes (number of drives from a particular vendor, number of readable drives by interface type as examples) may be low which in turn means there would be a higher margin of error for that particular comparison or evaluation. Whenever a margin of error of ± 20% was encountered (typically when a sample size was 24 or less, as per Table 3) it would be explicitly noted and discussed as such.

4.3.2.3 Research questions

As researchers may have differing questions they wish to answer it was not included in Table 2 on page 59 for comparisons. As the literature review showed however, typically "How many drives are wiped?" and "How many drives had information that could identify companies and/or individuals from the remnant data found?" are two measurable, comparable questions that can be answered for the research in this thesis. A potential research question arising from the mobile phone research is "What relationship is there between the cost of the mobile phone and the state of data on the phone?" was not considered practical when translated to hard drives. The ambitious question of "Which companies are more at risk from data remanence?" was considered as it had not been addressed adequately by previous researchers.

4.3.2.4 Identifying Information

As Table 2 shows not all of the prior methodologies make it clear exactly how they define "identifying information". Overall the consensus for "identifying information" however that is it is more than what merely appears in a telephone directory. In other words, finding an instance of a person's name and telephone number or address would not typically classify an entire hard drive as having "identifying information". One of the considerations raised in the literature review (Chapter 3) is that in certain countries such as the USA and Korea, citizens have a specific identifying number associated with them, and finding that number on a hard drive may go a long way to identifying that person and would be a useful piece of data. As Hooper and Evans (2010) state however there is currently no unique identifier for New Zealanders.

However by examining the drives that the previous researchers have discussed it is possible to reasonably assume what they considered to be identifying information. An example from Section 2.6.3.4 (on page 37); "On one disk the documents recovered revealed the user's name, occupation, the specific organisation and address at which they worked, as well as telephone numbers, email addresses, a complete employment history and details of their personal education. A large amount of personal email was also recovered" (Jones, et al., 2010, pg. 101).

In this thesis a "company" typically refers to a business, but it may also include not-for-profit organisations, educational institutions such as high-schools, and also government agencies (in Section 1.2 on page 2).

For this research a hard drive was classified as having individually identifying or company identifying information based on the following criteria:

1. For individuals: It is likely that the owner or one common user of the hard drive was identifiable from common details such as name/address/phone details but also from personal email correspondence or personal documents stored on the hard drive as per the example above.
2. For companies: Where multiple details are held on various individuals but may not be as conclusive as 1) above.
3. For companies: The company must be able to be classified according to the *ANZSIC* classifications (see Appendix H for the Level 1 classifications) other than the "T" classification.
4. For companies: Finding logos, business quotes, tax or other financial records and other related information.

The term "company" has been used partially to make it clear that it is distinct from businesses such as "sole traders" (one person running their own business) and for this thesis it also means "amateur sports clubs", "hobby groups" and the like are also excluded. This decision was made contrary to Kwon et al. (2006) who did list at least one amateur club; however the majority of the company examples given by other researchers were large to very large businesses.

Therefore a hard drive typically either had a large amount of specific data on one specific person, or large amounts of data on different people. In the second case the hard drive would also typically belong to a company (such as a high school) or have multiple users logging into that hard drive who worked for the company. There may be instances where the company is identifiable due to emails or other common references to the company's name but an individual user or owner of the hard drive may not be identifiable.

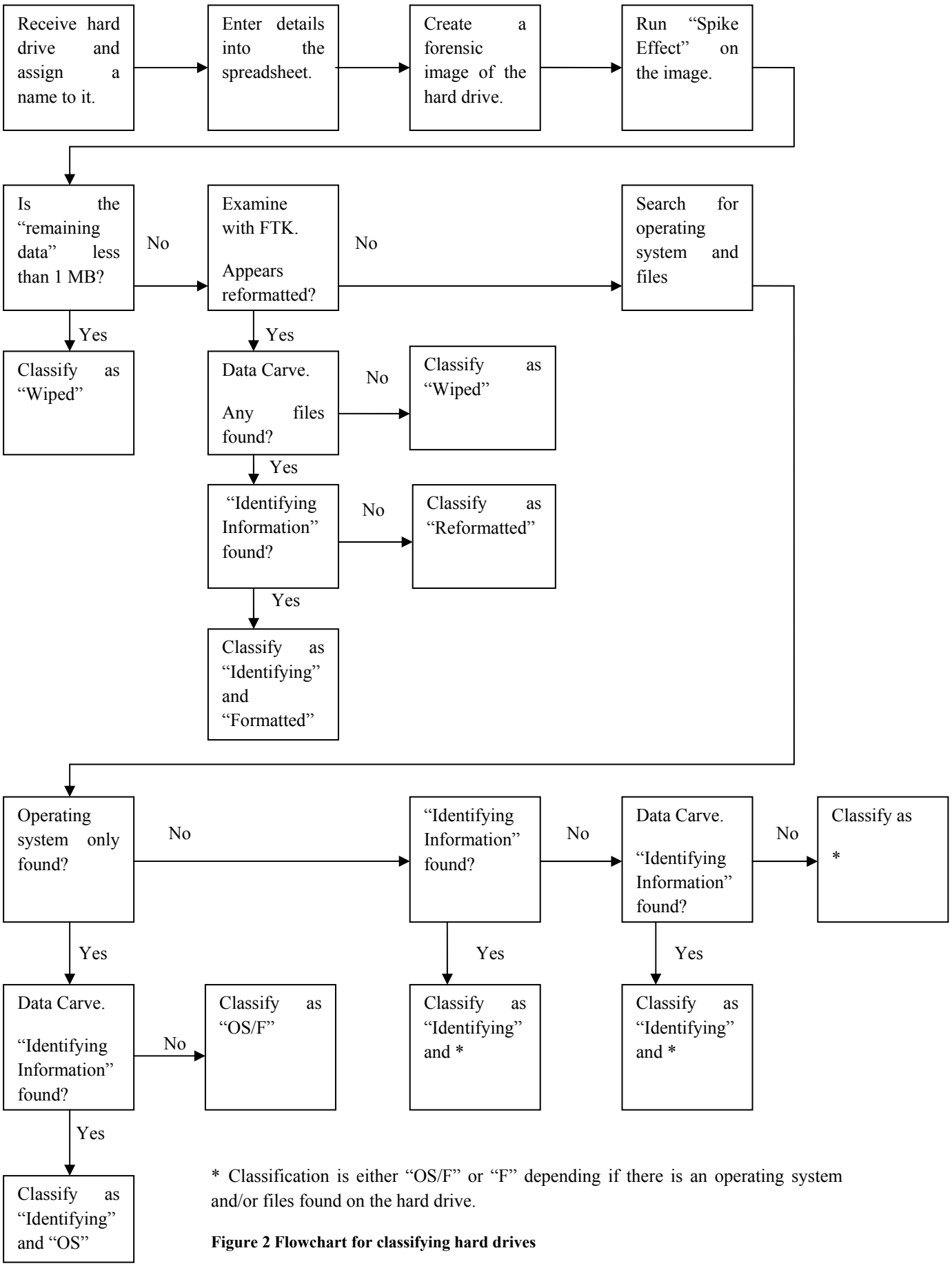
4.3.3 Experimental procedure

After considering the methodologies as they were described by other researchers, the research questions that were to be addressed and that creating an automation tool was possible, a methodology was designed that would be clear for other researchers to follow. The overall process for the experiments conducted is illustrated by the flowchart (see Figure 2 on page 65). The process uses terms such as "Spike Effect" which is described in the glossary and

also discussed in depth in Section 4.8 on page 77, *Forensics Toolkit (FTK)* which is was the forensics analysis tool of choice (discussed in Section 4.9.1 on page 92) and data carving.

Data carving was used as it allowed the results to be differentiated into information typical users who had purchased a hard drive on the second hand market could find, versus users who had an interest in finding out potential further information on any former owners of the second hand hard drives.

The main changes to the experimental procedure, when considering mobile phone data remanence research, is possibly changing the analysis tool (*FTK*), potentially changing the "remaining data threshold" from one megabyte and testing the Spike Effect program to ensure that it works with other data storage devices if applicable.



* Classification is either "OS/F" or "F" depending if there is an operating system and/or files found on the hard drive.

Figure 2 Flowchart for classifying hard drives

4.4 Documentation

For this research a number of documents were created. These include "post-it" notes placed on each hard drive when they arrived, and a spreadsheet for hard drive physical properties such as drive type. The naming convention for drives was a straightforward drive interface type (such as SCSI, IDE, SATA), and a sequence number. For example the first hard drive named was a SCSI drive so it was SCSI_01. The 48th drive was an IDE drive so it was IDE_48.

Each source of drives also had a naming convention. In that case it was "C" for private company or "A" for auction site, a letter (starting at A) to represent a specific company or auction, and a number to represent the shipment from that company. Auctions were based on the auction site's name. In this case only Trade Me was used, however the naming convention was designed to include other auction sites over time. An example is C_A_1 was the first shipment from Company A. C_B_1 represented the 1 and only shipment from Company B. A_T_1 represented the first Trade Me auction.

The post-it notes had a name field, and tick boxes for "Imaged", "Scrubbed", and "Unreadable". "Scrubbed" was used internally to denote that the author had wiped the drives in preparation for disposal, rather than using "wiped" which may have implied the drive had arrived already in a "wiped" state. This made it possible to tell at a glance which phase of the experiment the drive was in. As the final step for those drives that were being returned to source (drives donated by private companies), the post-it notes were removed, and in the case of "Unreadable" a new post-it note was put on the drive with "Unreadable" and a brief description if possible of what the problem encountered was.

The spreadsheet of physical drive properties had the following fields: a sequential number (starting at 1), the hard drive serial number (so that it would be possible to tell if a drive had been sent and processed before), the drive project identifier using the naming convention as described above, hard drive capacity in gigabytes and arrival date. Additional fields were filled in as the experiment proceeded. These were "Scrubbed", "Unreadable", date returned to source and "Notes". These fields aligned with the fields on the post-it notes and also ensured that all drives were accounted for. The difference between date arrived and date returned was used for time management. The hard drive creation date would have been a potentially useful piece of information as there may have been a connection between newer drives being sanitised due to the "SECURE DELETE" technology or owners of newer drives being more

concerned about data remanence issues. However not all drives had the date on their labels, and it was considered that the trade-off for looking for the hard drive creation date was not worth the information it might produce. The "Notes" field was primarily used to note anything about the drive that was unusual and in the case of SATA and IDE drives where *Tableau* (the hardware used to forensically copy the SATA and IDE drives) noted "The source drive may be blank".

A spreadsheet for what was contained on the drives was also created. This spreadsheet had three fields for clearly viewable drives (Operating system only, Files only, Operating system and files), if a drive had any identifying information, if the drive was reformatted, if the drive had been wiped and if the drive was unreadable.

Another spreadsheet was created for drives with identifying information specifically. This spreadsheet had four columns: hard drive name, if it was company identifying (Y/N), individually identifying (Y/N) and reformatted (Y/N). From those values it was straightforward to calculate how many drives had at least company identifying information and were reformatted for example. The overall purpose of this spreadsheet was to attempt to clarify the issue with the consortium reporting where it was unclear the total number of drives that had identifying information and which of those drives required additional tools to find identifying data.

4.5 Sources of Hard Drives

The primary objective of this research was to create a baseline of data remanence within New Zealand. As such, there were a number of considerations based on the previous research and other factors that may only relate to the New Zealand context. Considerations included but were not limited to:

- Replicating previous research
- Cost, both initial and ongoing
- Any potential bias from the source(s)
- Reliability of the supplier(s)
- Consistency of supply
- Timeframe for receiving the hard drives
- Other issues

Jones, et al. (2010, pg. 198) recommended that "disks that are obtained for the study are selected from the widest possible range of sources" but they did not elaborate on how many different sources they used. The 2009 consortium analysis (Jones et al., 2010) also noted that they excluded drives where the vendors claimed that the drivers were wiped. For this thesis where vendors have made explicit statements about the state of data on the hard drive (wiped or not wiped for example), that drive would be classified as "known state" drives. El Emam et al. (2007) did however have known state drives as they were able to compare "statements made by the vendor" versus what El Emam et al. found. This research uses a selection method that excluded "known state" drives from the selection process. The following sections consider the possible sources for second hand hard drives in New Zealand by source type and then by the above factors.

4.5.1 Auction sites

Due to New Zealand's geographic size and reasonably low population, physical auction houses were ruled out as a source of drives. This is because a search of the New Zealand yellow pages for auction houses in Dunedin returned 21 results (Yellow) however, they included car only auction businesses, real estate only auction businesses and other businesses unlikely to be selling second hand computers or hard drives. New Zealand has a number of online auction sites such as Trade Me, www.trademe.co.nz , Sella, www.sella.co.nz , and bid4it www.bid4it.co.nz (TradeIT, 2008). Trade Me is the largest of the sites and was the only auction site to be selected. That was due to overall reliability, reputation and number of lots on offer.

Replication of previous research: The consortium members reported they use auction sites as the primary source for their experiments. Examples of this include the eBay online auction site. A simple examination shows Trade Me is similar to eBay.

Cost, both initial and ongoing: With purchasing second hand hard drives from auction sites there were two main options. The first option was to source an appropriate number of drives (the total for the experiments), forensically capture them all, analyse them all, and then resell them, most likely back on auction sites, or if funding allowed it, donate them to a worthy cause. This would have required a reasonably large outlay of money. It may have also been difficult to budget exactly how much money was needed to purchase the bulk lots of drives, as bidding on many lots at once may force the prices up. The other issue would also be securely storing all the hard drives at one time.

The second option was to buy a small lot of drives, image, analyse, resell, and use the funds from the reselling to purchase the next set of drives. While this has the benefit of requiring less initial funding and require less physical storage space it could present problems if the methodology had to be revisited for capturing data and the drives were physically required.

The decision was made to purchase all the hard drives, and store them until the experiments were finished. While this required a greater outlay it also meant that in the unlikely event the drives need to be re-examined they were directly available.

Any potential bias from the source(s): The drives were selected where they were not "known state" that is the vendor did not make a statement if any data remained on the drives or not. Selecting small lots from different vendors on an auction site lowers any potential bias in the source.

Reliability of the supplier(s): As auction sites provide many different potential suppliers, then there should be no real concerns with the reliability of supplier. In the cases of fraud or missing shipments there are policies and procedures in place to deal with this, such as Trade Me's recommendations and procedures. (Trade Me, 2011). That guide outlines waiting seven days for the seller to act, and then contacting Trade Me who will then resolve the issue.

Consistency of supply: As with the reliability of supplier, given the large volume of potential suppliers then it would be expected there would be a consistent supply of second hand hard drives to conduct research on.

Timeframe for receiving the hard drives: Under the first purchasing option of buying the entire consignment of drives at once, it would be expected to take a week or so for all the drives to arrive. Under the second option, there would be a time delay associated with reselling each batch of drives, receiving payment and then bidding on replacement drives.

Other Issues: If the second purchasing option was taken (buy a small lot, process, resell on the same auction site), it was anticipated there could be a small chance that drives that had been previously purchased would be purchased again. While this is probably unlikely to happen, it would add additional costs and processing time. This risk is partially mitigated by recording the unique serial number of each hard drive and comparing serial numbers before adding a hard drive to the spreadsheet.

4.5.2 Private companies that sell second hand hard drives

Replicating other experiments: As mentioned in the previous section, the consortium experiments were predominantly sourced from auction sites, so sourcing hard drives from private companies that sell second hand hard drives may produce different results. This is a consideration as companies may have internal policies to sanitise all drives before reselling them (but do not mention that to buyers).

Cost, both initial and ongoing: Depending on the arrangements made, the private companies may be willing to donate drives as part of public relations, aiding research, or for the service of having those drives properly forensically wiped. Alternatively the private companies may sell the drives at a specific rate, and then the drives would have to be disposed of, either via on-selling or being donated.

Potential bias: Private companies may be able to buy large bulk lots from schools or government agencies and therefore drives purchased from private companies may have large runs of similar results. This can be mitigated by making smaller purchases over time, or by going to the private company in person and supervising the selection process which may or may not be practical and may not be permitted via company policy for security reasons.

Reliability of the supplier(s): There is always a risk that a supplier may not be able to continue to supply drives after the research has started. Having a number of different private suppliers would lower that risk. If a formal agreement has been signed this too may mitigate that risk.

Consistency of supply: A possible concern with dealing with only one private company supplier is that they may have sourced their drives from one specific area/organisation type such as only Wellington or only Auckland, or only government agencies, or only high schools, or only hospitals.

Timeframe for receiving the hard drives: This would depend on the arrangements made, and how much time the private company wished to devote to the research. As stated above, a formal agreement could help define the timeframe.

4.6 Acquiring the data

As stated in Section 2.4.2 (pages 25 and 26) there are two main ways to acquire data in data remanence research, by using dead forensics or by using live forensics. For this research dead forensics was used as it meant the original source was not altered during the capture process or during the analysis phase. The dead forensics approach appeared to be the preferred method used by other researchers for similar reasons with the exception of Medlin and Cazier (2011) who stated that they were conducting their research in such a way that it would be similar to a typical user.

There were two main factors regarding acquisition of the data. Those factors were the hard drive interface and whether the drive was readable or not.

4.6.1 Hardware used

For SATA and IDE hard drives a *Tableau* was used as it is specifically designed and marketed as a forensics duplicator. The *Tableau* has been examined by NIST. The vendor description of the *Tableau* is included in Appendix G. Briefly, it is a forensic duplicator that can duplicate at up to 6 GB/minute, uses MD5 and SHA-1 hash functions, detects hidden disk areas and a "blank checking feature" (Tableau LLC, 2009, pg. 4).

One of the benefits of the *Tableau* therefore was the speed at which it could bitwise copy a source hard drive (the hard drive that had been purchased second hand) and also the speed it wrote files to the destination hard drive. During the experiments in most cases it met the claims of up to 6 GB/minute (lower rates were encountered due to limits of some hard drive interfaces. This differs from the consortium experiments as the *Tableau* is a stand-alone tool, rather than being a piece of software installed on a standard desktop.

One of the main features that the *Tableau* also provides is a "blank check" feature. This feature allows the *Tableau* to quickly determine if a destination drive is blank (important to ensure previous case files are not accidently overwritten and that there is enough space to write a new case file to) and to potentially speed up forensics analysis of a source drive image. The *Tableau* manual however states the following with regards to "source drive may be blank" (Tableau LLC, 2009, pg. 18):

> "The TD1 checks selected sectors on the source disk looking for non-blank data patterns. If all of the checked sectors appear to be blank, the TD1 warns the user that the source *may* be blank. This does not mean that the source *is* blank, but it might

mean that the source has been partially wiped or that an ATA password has been set for the source drive."

Furthermore, a *Tableau* user may initiate a "Quick Check" themselves, as described by the *Tableau* manual (Tableau LLC, 2009, pg. 36):

> "When performing a "Quick Check" the TD1 reads sectors in the *Master Boot Record*, the *Primary GPT*, and the *Secondary GPT*. A sector is considered to be blank if it contains only a repeating pattern such as 00h, E5h, or FFh. Any non-repeating pattern is considered to be non-blank. If all sectors read by the TD1 have repeating patterns (though not necessarily the same repeating pattern), then the TD1 concludes the drive may be blank.
>
> ***Important: A "Quick Check" is not an exhaustive check of the entire drive. It is possible for a drive to appear to be blank according to the quick check while still storing forensically relevant information. A forensic examiner should treat "blank" source disks with some suspicion and use other tools, like a Tableau write blocker, to examine the drive to see if it contains forensically relevant information.***"

The terms "Master Boot Record" (MBR), Primary GPT (Guid Partition Table) and Secondary GPT (Guid Partition Table) are discussed briefly in Section 2.2.3.2 and 2.2.3.3 (pages 16 and 17). Of note, the three hexadecimal codes it looks for as stated above are 00h, E5h and FFh which translate to 0 decimal, 229 decimal and 255 decimal. It is therefore expected if *Tableau* thinks a source drive is blank, that the Spike Effect would find a large spike or spikes on those three values.

For SCSI hard drives a normal computer was modified by inserting a SCSI card, and the *FTK Imager* software was used. The *FTK Imager* is conceptually very similar to the *Tableau* device in that it allows for forensic imaging of hard drives and creating case files however it is implemented as software. The default settings were used, and this meant that only one case file was made per hard drive captured. Similar to *Tableau*, *FTK Imager* uses hashing to determine if the capture is identical to the source drive.

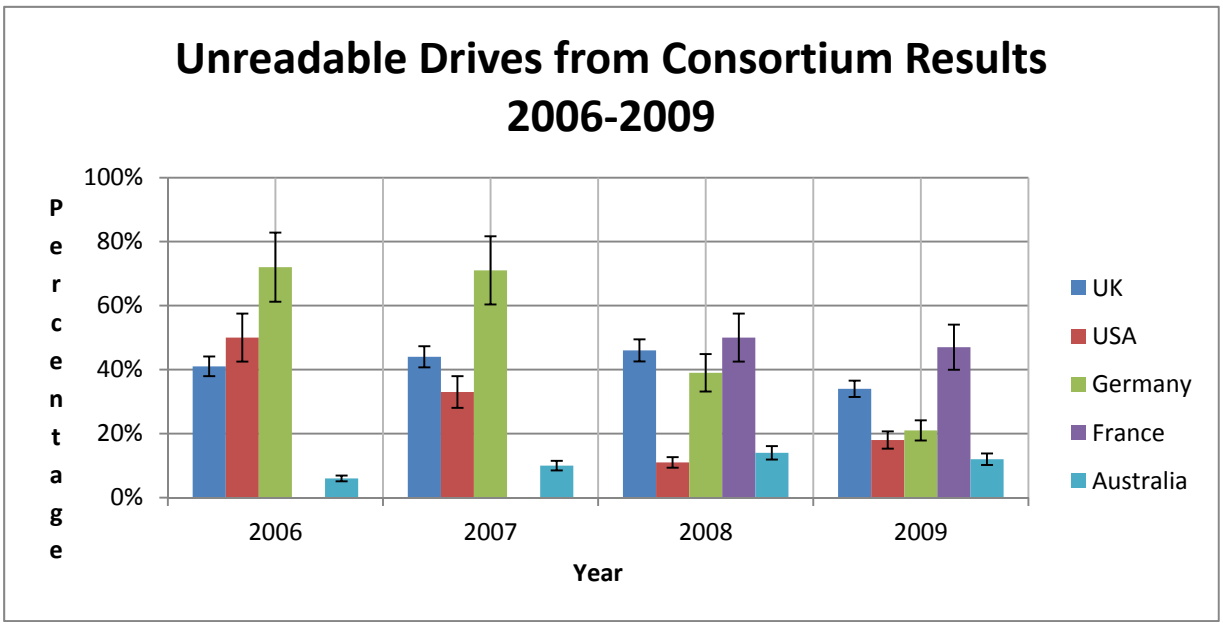
Both capture solutions were tested with "known state" drives before attempting to capture data from the hard drives used in the experiments. This initial testing involved forensically wiping the hard drives, capturing the hard drives and examining the data found. Known data

were then put on the drives, the process repeated and finally the drives were wiped again and the process repeated again. This ensured that the process of capturing did not modify the data in any way and that the data captured matched the known data put onto the drives. There were no issues found in the testing so it was expected that the process would work with unknown state drives.

4.6.2 Determining whether a drive was unreadable

For this thesis the author considered the potential issue of unreadable drives and further examining the consortium results enabled the creation of Table 4. Table 4 below considers the consortium results for 2006-2009 when considering the percentage of unreadable drives in a sample by country. The values are based on Tables 28-31 in Appendix I. The error bars were set at 7.5% for the United Kingdom as their samples were approximately 171 drives per year or more, and the error bars for other countries were set at 15% as sample sizes ranged from 24 to 96. This is an approximation for consistency and was discussed in Section 4.3.2.2 on page 60. Additionally, values for France were not reported before 2008.

Table 4 Comparison of Unreadable Drives from Consortium Results 2006-2009



As Table 4 does however show, the percentage of unreadable drives can range from 8% to 80% but after considering the outliers is closer to 20% to 50%. While Sutherland et al. (2010) outline potential solutions, they do not outline the likely timeframe to make such repairs or all the resources and skills needed, so the decision was made that changing the

jumper settings and re-attempting to forensically capture the hard drive would be an acceptable solution. This is because the jumpers are external on hard drives, are often well documented (usually with diagrams on the hard drives) and would take a very short time to change.

In certain cases, lights and sounds or lack thereof, or error messages displayed onscreen would immediately alert the author that the drive could not be captured. The best example of the lack of sound issue was when a hard drive was expected to spin up which usually produces a whirring sound. In at least one case, the drive would not spin up and this suggests there were issues with the power cables or power supply. As the drive could not be powered on, it could not easily be read. Examples of error messages included "Start unit request failed" which was noted on a SCSI drive. This error message is described as resulting from a number of possible causes including bad sectors or more specifically from the hard drive having been powered up properly, but not powered down properly. In the case of SATA or IDE drives error messages including bad sectors would be displayed on the *Tableau*'s screen.

The computer used for capturing data from SCSI drives would normally take 30-40 seconds to go from the Windows XP splash screen to the login prompt screen. If there were problems with the SCSI drive being read that were not stated on screen, the Windows XP splash screen would typically be displayed for at least three minutes.

In those cases the time would be noted, and the experiment would be allowed to continue for three minutes. If there was no change to the system at that point, it would be powered down, the hard drive and cabling would be checked and changed if applicable and then the experiment would be restarted. If it reached the same state, it would be powered down again, and a known good hard drive would be used to check that the system itself was working. Once the system was verified as good, a third and final attempt to capture the drive would be made. If it reached the same "suspected error" state as before, the drive would then be recorded as "unreadable".

In cases where the hard drive was readable, the forensics capture of the data was named and stored on either the capturing computer in the case of SCSI drives, or on a specific hard drive for IDE and SATA drives. The files associated with the captured data would then be stored on an external USB one terabyte drive as a backup.

4.7 States of data on readable drives

This section describes the different states of the data on the hard drives such as "wiped", "reformatted" or "clearly viewable". The different states have also been listed in order of least recoverable and/or viewable to most recoverable/viewable.

Each section gives a simple description of the state followed by a more in depth discussion. In instances where the files are clearly viewable (or in easily recoverable so that they become clearly viewable such as "reformatted" or "deleted") the different types of files have been considered.

4.7.1 Wiped

This means that any data that were on the hard drive was not recoverable with the tools available to the author. As outlined by the "Source may be blank" check in *Tableau*, this typically means that the hard drive data has been overwritten with a specific byte or bytes such as NULL. This classification is a subset of the sanitisation procedures listed in publications by both NIST(NIST, 2006) and by New Zealand Government Communications Security Bureau (GCSB). The guidelines from NIST are included in Appendix F.

The New Zealand guidelines are somewhat similar as NIST with regards to clearing data (New Zealand Government Communications Security Bureau, 2010, pg. 156):

> "**Non-volatile magnetic media sanitisation**
>
> ***System Classification(s): R, C, S, TS; Compliance: must***
>
> Agencies must sanitise non-volatile magnetic media by:
>
> • if pre-2001 or under 15GB: overwriting the media at least three times in its entirety with an arbitrary pattern followed by a read back for verification, or
>
> • if post-2001 or over 15GB: overwriting the media at least once in its entirety with an arbitrary pattern followed by a read back for verification."

Therefore both documents make it reasonably clear how media such as hard drives should be handled. The GCSB documents are usually aimed at government agencies that may have "top secret" data, but it can also be used by non-government agencies such as businesses and schools.

For the purposes of this research the author is concerned with determining whether the former owners have been merely disposing of their second hand hard drives or whether they have

been making attempts to clear those drives. If this research had included microscopy then purged (as described in Appendix F) drives would also be a consideration.

The tools used for the acquisition and clearing of the hard drives in this research describe their clearing functions as "wiping". Those tools allow for a single write or a multi-write overwrite of data and in the case of *Encase* the user can specify the pattern to be used.

For the purposes of this research the terms "wiped" and "wiping" were used to describe drives where an arbitrary pattern was detected and little to no other meaningful data were found.

As analysis with the Spike Effect showed (Section 4.8 on page 80) certain drives did have a very small amount of remaining data on the drive. For practical purposes these drives have been classified as wiped, as the data found were in the magnitude of kilobytes on drives that had gigabytes of capacity. In other words the data found were less than one millionth of the drive capacity, and was thus not likely to be useful in any meaningful way.

4.7.2 Reformatted

This means that the operating system knows to reuse the space occupied by the data and does not display these data to the user. When a drive is reformatted with the Windows Operating system files are typically marked with a special code at the beginning of the file name. Typically that code is E5 hex for Windows operating systems (which is one of the codes that *Tableau* uses to determine if the source drive may be blank). This means that data are still on the hard drive but are flagged to be used by the operating system and not displayed to the typical user without using specialised tools. The consortium use the term "formatted" to describe drives where attempts have been made to remove remaining data either via reformatting/repartitioning a drive and/or installing a new operating system onto the drive. This thesis uses the term "reformatting" to mean the same thing as it is slightly clearer that the user has taken an action to remove data.

4.7.3 Clearly viewable

This means that the data can be easily seen when the hard drive is turned on and would therefore be easily found by a typical user. This is the approach that Medlin and Cazier (2011) took. Previous researchers have included tables of commonly found data such as Garfinkel and Shelat's (2003, pg. 25) table outlining occurrence rates of common MS Office files such as Word .doc and PowerPoint .ppt. Medlin and Cazier (2011, pg. 32) include a

table of viewable data such as Word documents containing banking information or other financial information (in their research they also found a person's will), and image files such as JPEGs containing photographs or diagrams.

This author had initially intended to record the counts of file types found however since some hard drives had over 400 000 files whereas other drives had less than 5000, this meant comparisons or aggregates would have been potentially of little use.

4.7.4 Special case: Encrypted

As outlined by Thomas and Tryfonas (2007) one of the ways of protecting data on a hard drive is to encrypt them. Hard drives can range from those with only small amounts of data that has been encrypted such as personal or business emails to "full drive encryption". Thomas and Tryfonas assert that (2007, pg. 465):

> "Individuals and organisations should also consider the full encryption of hard disks so that if the disk is lost or the data is not effectively removed on disposal, it will not be easily recoverable. This would provide adequate protection in most situations...The new Windows Vista operating system has improved support for data protection at all levels with a full volume encryption of the system volume, including Windows system files and the hibernation file, which helps protect data from being compromised on a lost or stolen machine."

Joukov et al. (2006) considered a number of myths regarding secure deletion and the myth that encryption can be an effective form of data "deletion" (since the data are not retrievable in readable form). Joukov et al. explain the potential problems associated with attempting to use encryption for that purpose. The author took the view that, much like unreadable hard drives, if any fully encrypted hard drives were found, no attempts would be made to decrypt those hard drives. The drive would be classified as "wiped". The main reason for this classification is that the time taken would likely range into lifetimes of computing power required for a very limited outcome.

4.8 Automating the research using the Spike Effect

The Spike Effect is the name of both a program and the technique developed by the author for analysing evidence files. Before explaining the Spike Effect fully it is important to understand that 1 gigabyte represents (approximately) 1 billion bytes and therefore 1% would

represent 10 megabytes or 10 million bytes. Because so much information can be stored in 10 megabytes, this is not enough of a granular measure, percentages in this section are often rounded and noted as such. Where the percentages are not rounded they are expressed to at least 4 decimal places.

Based on the model by Fragkos et al. (2006) for automating forensic analysis, rejecting drives that are unreadable is trivial as the tools used to capture the data from those drives will either give an error message, or the drives themselves will not start which alerts the researcher conducting the capture that the drive is unreadable.

Determining whether a drive is wiped is a different process. Typically it would either require using tools such as *Encase*, *FTK* or similar to examine the actual data in the evidence files, or using search phrases to conduct a search.

The tools used to forensically capture the source hard drives would create a number of evidence files (called a case) in a specific directory. Evidence files created by the *Tableau* shared the same base name, and have an incremental counter as the extension such as thesis.001, thesis.002, thesis.003. Those evidence files were each approximately 4 GB in size so a 20 GB hard drive would generate five evidence files of approximately 4 GB and a sixth file that was the remainder. When the SCSI drives were captured, *FTK Imager* was used and only one evidence file (the same size of the hard drive's capacity) was generated per case. The author did not select any compressing options with the tools when capturing the evidence files.

It was considered it may be easier to write a very simple piece of software (that calculated byte histograms on very large files) to check a directory for a specified case and analyse the case files rather than loading each case into *Encase* or *FTK* for analysis.

The author wrote a program using standard C so the program could be easily ported to other operating systems such as embedded systems. One of the considerations for the Spike Effect program was that a 32 bit unsigned integer (commonly referred to as "INT") ranges from 0 to 4,294,967,295. As some case files or cases may be larger than 4 GB (such as the SCSI_03 case file being one 18 GB file) then counts run the risk of over overflowing and producing incorrect values. To avoid this, unsigned Int64 was used which ranges from 0 to 18,446,744,073,709,551,615. This also required using the 64 bit functions instead of 32 bit functions. The program opens the case files, reads in all the bytes and then analyses them.

One of the outputs of the program is the total number of bytes read in. This is compared with the size of the evidence files, which in turn is compared with the size of the hard drive. Any major disparities would then require further investigation. Those disparities could include a drive listed as 18 GB generating an evidence file that was only 16 GB. That would suggest some sort of hidden disk area (Section 2.2.4.1), or that the case files were not a true bitwise copy of the source drive.

The program was automated in two ways. First when started it would search the entire directory selected for any matching case files (as only one case should be in a directory this was not likely to be a problem and did not end up posing a problem), and secondly a batch file was used to enable the program to be run multiple times, with different case directories specified each time. In practice the program was run overnight to generate the output files for multiple cases and worked well in that capacity.

The performance of the Spike Effect program was compared with two other freeware byte histogram generators; *JByteStat* (version 2.0) and *Bytehist* (version 1.0 beta 1). Both of those programs state that they can read files of any size, and generate appropriate histograms based on the single byte values found.

Both programs use a graphic user interface (GUI) for file selection and presenting the results which the Spike Effect program does not currently do. *Bytehist* allows for the histograms to be saved as image files such as JPEG and also has a non-gui approach for scripting purposes. Applying the scripting functionality to a case such as IDE_71 (An 80 GB drive with 21 evidence files) *Bytehist* would treat each evidence file independently and generate a histogram output image for each file (resulting in 21 image files), rather than the Spike Effect program which reads in each evidence file and generates one text file for the entire case. *JByteStat* did not appear to have any output options beyond the user performing a screen capture such as "print screen" themselves and did not allow for analysing more than one evidence file at a time or in a scripted manner.

Testing *JByteStat* and *Bytehist* with specific test files (4 Kb, 250 MB, 4 GB) showed the programs ran at approximately the same speed as the Spike Effect program. *Byehist* was not able to process the 4 GB test file on the systems used for testing. The time taken by *JByteStat* and *Bytehist* was based on noting the start and stop times when the programs were independently run, as neither appeared to have an in-built function to display the total processing time. The Spike Effect program uses two timers, one for the overall program

running time, and one for the processing time of each evidence file. Neither *JByteStat* nor *Bytehist* output the histograms into text files for further processing or examination which the Spike Effect program does.

The output files from the Spike Effect program were then entered into an Excel spreadsheet that in turn calculated and displayed the values from the case files. Initially only the largest value or "spike" in the data were measured and accounted for in the size of the drive. In other words if the null byte (0) occurred 810 000 times on a one megabyte drive it would be displayed as the largest spike. The spreadsheet calculates the percentage of each value (81% in the above example), and how much drive space was left over after the largest value was accounted for (290 kilobytes in this example). As will be demonstrated however, measuring one spike only was not always a reliable indicator of the hard drive's nature.

4.8.1 Results of running the Spike Effect program over specific drives

The results of running the Spike Effect program over the case files produced two main types of result that were then classified as ("Totally Wiped" or "Very Diverse") and two sub-classifications ("Practically Totally Wiped" or "Special Case"). Each classification is explained with an example of the drive that produced those results. The results have been graphed using a column graph to show the "spikes" in the data.

"Totally Wiped" is a category where is only one spike and therefore there is no possible remaining data. Drive IDE_44 was a 20 GB drive. It had approximately 20 billion entries in the Hexadecimal (Hex) 97 / Decimal (Dec) 151 array field and 0 entries in each other field which meant the array entry "151" comprised 100% of the array entries (see Figure 3 on the next page). Therefore there was no remaining space.

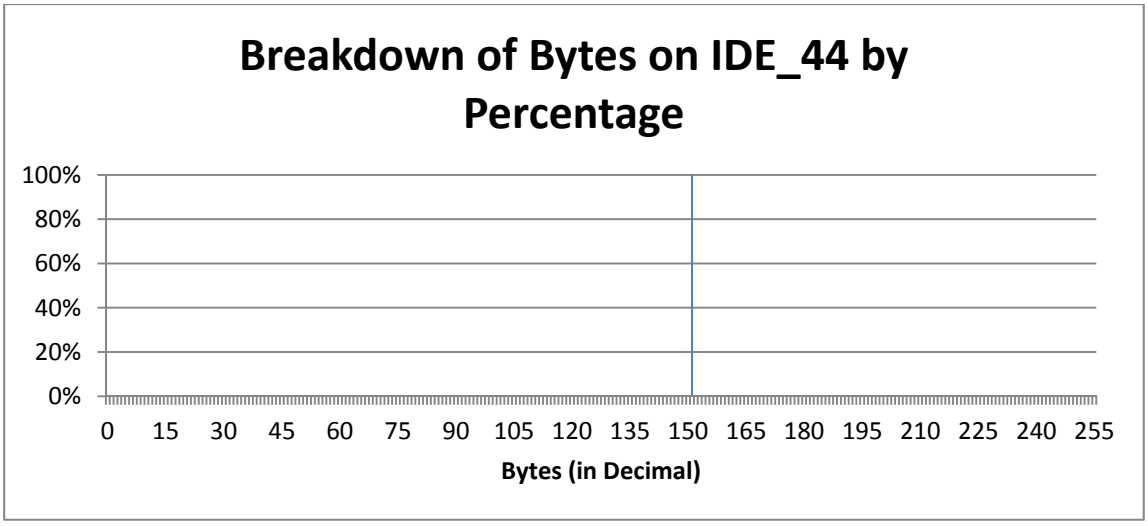


Figure 3 Example of "Totally wiped" drive

A subcategory ("Practically Wiped") was then developed for drives where after accounting for the three spikes with the most common values, there was one megabyte or less of remaining data. It was deemed that it was unlikely there would be much recoverable data in that remaining megabyte and drives that met the criteria of "Practically Totally Wiped" would not be examined further in *FTK*.

Drive IDE_39 was a 20 GB drive that had 59% of the drive written with Hex 97 / Dec 151, 40% as Hex CA / Dec 202 and Hex FF / Dec 255 making up the most common value of the remaining 1% (See Figure 4 below). It should be noted those numbers are rounded; the actual percentage occupied by the three largest values was 99.9999%. The remaining space in this case was 2,072 bytes or 2 kilobytes.

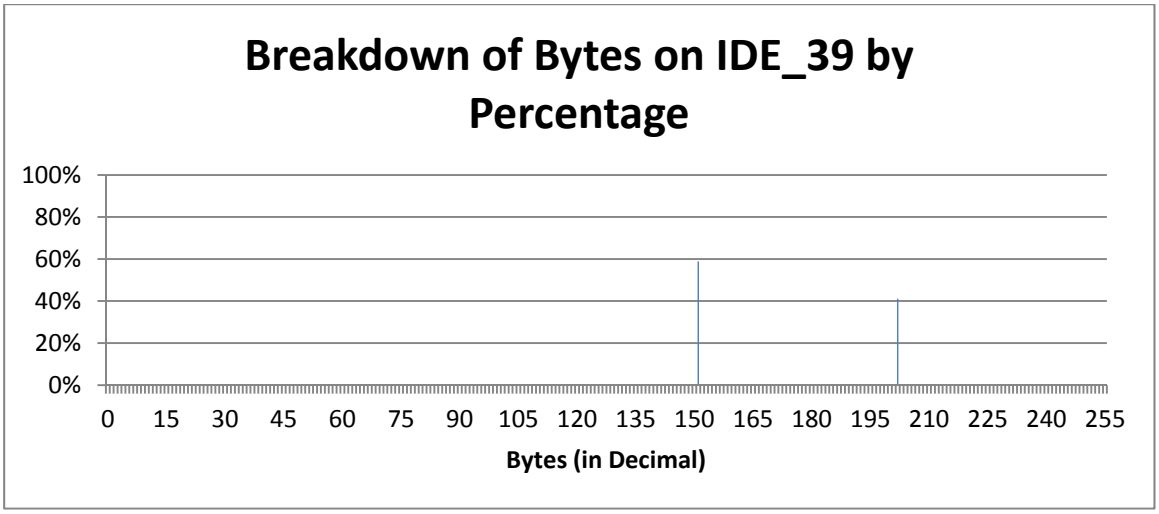


Figure 4 Example of "Practically wiped" drive

The opposite of a totally wiped drive therefore is a drive where all array entries have values in them. Typically those drives would have a fairly uniform distribution of values either before or after the three most common spikes were accounted for. The qualifier "very" has been used to denote a fairly uniform distribution and to allow for any sub-categories that may be found in the "diverse data" range.

Drive IDE_46 was 6.4 GB drive, and the three most common values only accounted for a total of 1.4% which left 6.3 GB of data on the drive. The drive contained all 256 possible values in a fairly uniform distribution (see Figure 5 below) so it is very likely that the drive has data remaining on it. The main observations are that there is a single spike at "0" of almost 0.6% and that the remaining values are all approximately 0.39%.

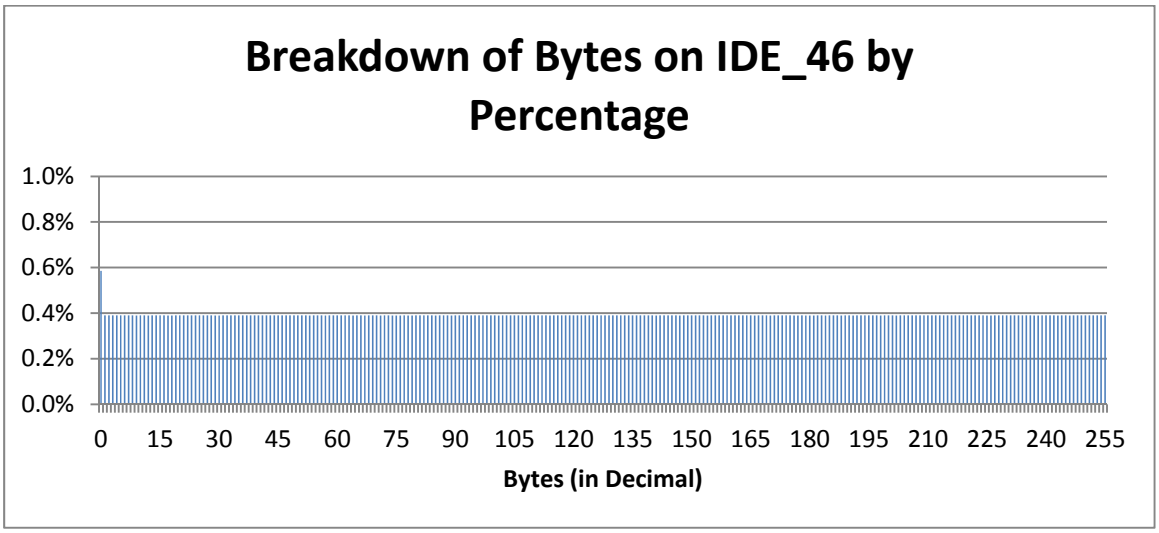


Figure 5 Example of "Very diverse data" drive

Finally there were a number of hard drives that did not meet the above three classifications and are currently classified as "special case". One such drive was SCSI_03 which was not "diverse" because it only had 16 spikes of any note but it did not meet the criteria of being "practically wiped" because it still had more than one megabyte of data remaining after the first three spikes were accounted for.

"Special case": SCSI_03 was an 18 GB drive. The three largest values accounted for 27% of the drive as three spikes of 9% each. After graphing the data (see Figure 6 on the next page) there appeared a pattern of 6 tall spikes of approximately 9% each, and 10 smaller spikes of 4.5% each.

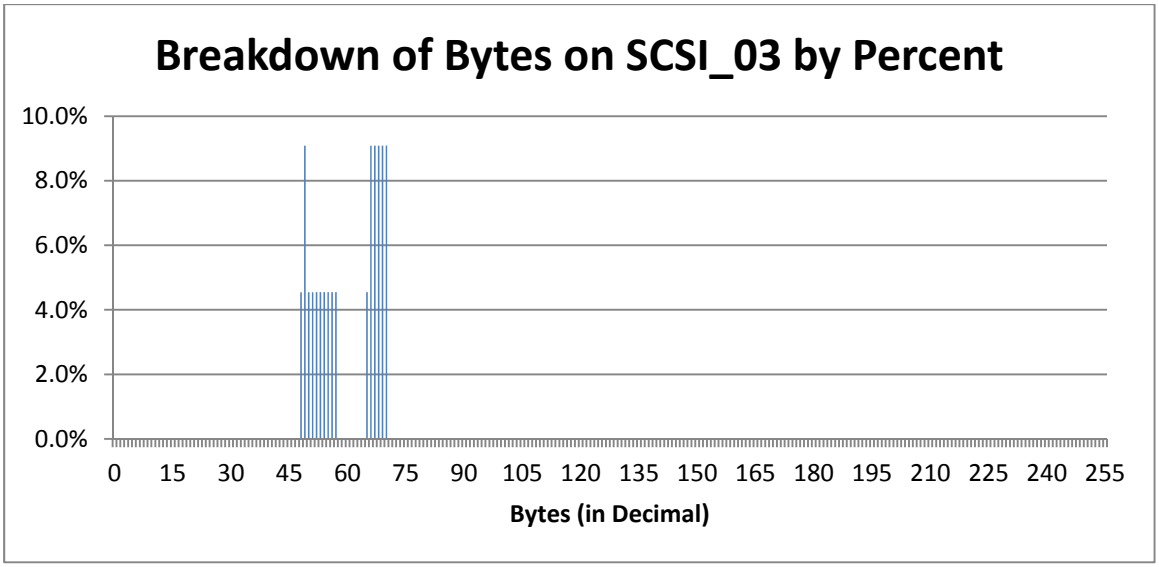


Figure 6 Example of a "Special Case" drive

Further analysis of the drive in *FTK* concurred with the most likely reason that a specific pattern had been used to wipe the drive. The Master Boot Record (MBR) from SCSI_03 was recovered and is presented in Figure 7.

```
Invalid partition tableError loading operating system


ThinkVantage Secure Data Disposal Utility 3.


Date and time of execution ... 11/14/06 16:04:43


Command executed ............. A:\BIN\SCRUB3.EXE /D=ALL
/W=1 /P=1F9E8D7C6B5A4F3E2D1C0B


Return code ................... 0 û`ò€þ??·Uª
```

Figure 7 Master Boot Record from SCSI_03

Characters which can cause formatting issues or which are "non-printable" have been removed from Figure 7 and a box has been placed around the text to make it clearer to read. The MBR showed that the drive had been wiped with a repeating pattern of 1F9E8D7C6B5A4F3E2D1C0B. Once all those data were accounted for there was virtually nothing left. There is also an 8000 byte entry for value 246 which is too small to be visible in Figure 6 on the previous page. Examining the pattern 1F9E8D7C6B5A4F3E2D1C0B showed two occurrences each of the characters B, C, D, E, F, and 1 which would explain six large spikes, and one occurrence each of A, 2, 3, 4, 5, 6, 7, 8, 9, and 0 which would explain the 10 smaller spikes.

The MBR provided an accurate explanation for the data found on the drive. It stated which wiping program had been used and how (Scrub3.exe), it also stated when the program was used (the date of "11/14/06"). The flag /P allowed for a pattern to be used when wiping and the pattern of 1F9E8D7C6B5A4F3E2D1C0B is consistent with what was found by the Spike Effect and by using *FTK*.

Figure 8 then compares the percentage of space remaining after three spikes has been accounted for on each of the four drives described above.

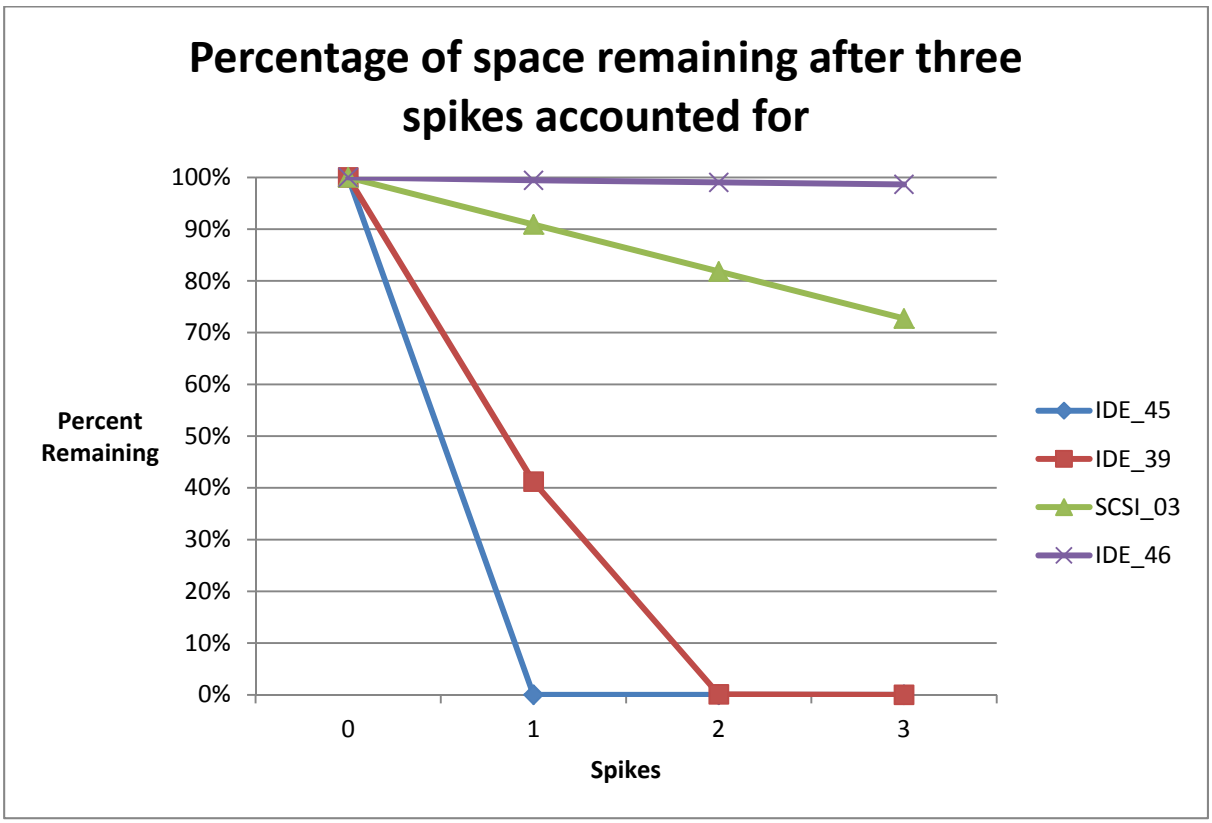


Figure 8 Comparison of the four drive types

4.8.2 A heuristic to determine whether a drive has been wiped or not.

Ideally to be fully automated the program would be able to determine for itself whether a drive was wiped or should be further examined by a tool such as *FTK*. There appear to be two main factors in determining whether a drive has been wiped. Those are the number of spikes, and the amount of data remaining after those spikes have been accounted for (see Figure 8 above). In the case of one spike for 100% (Drive IDE_44 in Figure 8) the drive can only be classified as wiped. Drive IDE_39 can be classified as practically wiped since there is so little data left after accounting for the first three spikes. If a drive has less than all 256 spikes with data in them then current results suggest that the drive will not have usable data on it.

There is one other consideration for the "Spike Effect" analysis. That would be to wipe a drive using a program that generated approximately uniformly distributed amounts of each of the 256 byte values. The "Spike Effect" would then report back a similar result to Drive IDE_46 as described on page 82 that the drive had readable data on it leading to a false positive. Other approaches that lower the distribution of different bytes used (a drive might only use 100 different bytes for example) but still maintains readable data on a hard drive would likely lead the Spike Effect to classify a drive as being a "special pattern". Whether the data are actually readable or not would need to be tested as no drives like that were encountered in this research.

4.8.3 Comparing the Spike Effect program with *FTK*

Firstly the processing times of the Spike Effect program was measured, and then compared with *FTK*.

Table 5 shows an example of Spike Effect program processing case IDE_44 which shows a reasonable consistent time of 2 minutes 42 seconds per 4 GB (As Table 5 shows the actual size of the case files are 3,999,989,760 bytes which is functionally close enough to be called 4 GB for the Spike Effect program).

Table 5 Example of the Spike Effect program processing time for IDE_44

File name	Time in seconds	Time in minutes	File size in bytes
IDE_44.ab.001	162	2m 42s	3,999,989,760
IDE_44.ab.002	174	2m 54s	3,999,989,760
IDE_44.ab.003	162	2m 42s	3,9999,897,60
IDE_44.ab.004	163	2m 43s	3,999,989,760
IDE_44.ab.005	162	2m 42s	3,999,989,760
IDE_44.ab.006	1	0m 1s	20,447,232
Total	**824**	**13m 44s**	**20,020,396,032**

There is a slight anomaly in the processing of IDE_44.ab.002 (approximately 12 seconds longer than the other files of the same size or approximately 7.4% longer) which was not fully investigated further as it only added 1.5% processing time to the entire case. The likely reasons considered were that depending when the Spike Effect program ran (i.e. directly after a reboot and login versus a logged in machine being idle for awhile) there may have been processes running such as the anti-virus doing a scheduled check for updates (the machine was not connected to any network so the system would not change), that there may have been issues with paging in and out of large amounts of data (reading a 4 GB evidence file, storing the array counts) and the like. As a note, Table 6 below shows a reasonably consistent 2 minutes 44 seconds per 4 GB case file across all the cases. As the Spike Effect program was at the proof of concept stage and run overnight, it was considered it would take longer to accurately determine the causes than the time saved (if any) by fixing those causes for these experiments.

Table 6 compares the Spike Effect time taken, the *FTK* time taken and the respective classifications. The *FTK* time does not include any additional time required for data carving. The hard drive classifications are based on the flowchart in Figure 2 on page 65. Times are in hour, minute, second (hh:mm:ss) format.

Table 6 Comparison of Spike Effect program and *FTK*

Case Name	**Case Size Gigabytes**	**Spike Effect Time Taken**	**Spike Effect Classification**	***FTK* Time Taken**	**Hard Drive Classification**
SCSI_02	18	0:12:29	Very diverse	1:30:14	Reformatted
SCSI_03	18	0:12:47	Special case	0:10:39	Wiped
SCSI_05	18	0:12:37	Very diverse	0:17:52	Reformatted
IDE_39	20	0:13:37	Practically wiped	0:06:42	Wiped
IDE_44	20	0:13:44	Wiped	0:07:53	Wiped
IDE_46	6.4	0:04:25	Very diverse	0:04:00	Reformatted
IDE_53	40	0:27:00	Very diverse	1:47:34	Identifying and OS/F
IDE_71	80	0:54:36	Very diverse	4:26:28	Identifying and OS/F
IDE_72	80	0:54:53	Wiped	0:34:41	Wiped

There are two main observations. The first observation where a drive was wiped with one byte only (IDE_44) or classified as wiped due to all three spikes adding up to 100% (IDE_39, IDE_72), *FTK* processed those drives faster than the Spike Effect program. The second observation is the disparity in time taken by *FTK* for drives of the same size that were classified as "Reformatted" such as SCSI_02 and SCSI_05. In that case the disparity is *FTK* took 1 hour 30 minutes and 14 seconds for SCSI_02 and only 17 minutes and 52 seconds for SCSI_05.

The first observation suggests that the Spike Effect program can be made much more efficient. One solution would be to have the Spike Effect program consider a sample of the bytes in a case rather than the entire case, such as every thousandth byte rather than every byte.

4.8.4 Developing a Spike Effect program variant that uses sampling

Consulting an online sample size calculator such as http://www.surveysystem.com/sscalc.htm showed that sampling every thousandth byte would produce a margin of error of ±0.05% when examining four million bytes out of four billion bytes (the size of a 4 gigabyte file).

Two variants were developed to test the effectiveness of sampling and the only change was from reading each byte to reading a subset of those bytes. This was to test if the variants were faster than the initial Spike Effect program, and also faster than *FTK*. Those variants were tested on the four drives examined in Figures 3-6 (IDE_39, 44, 46 and SCSI_03) as well as drives IDE_71 and IDE_72. Those two drives were included as those drives are the largest examined and are examples of "Very diverse" and "Wiped" respectively.

4.8.4.1 Systematic Sampling

The first variant used a systematic sampling approach. This means a *k* value is determined and every *k* bytes in the case files are analysed. As an example if *k* was selected to be 1000, byte 1, byte 1001, byte 2001, byte 3001 and so on would be analysed. To be a true systematic sample, the first byte must be selected from 1 to *k* inclusive. So in one case the first byte to be examined might be byte 35 and then 1035, 2035, 3035 and so on would be examined, and in another case the first byte to be examined might be 613, 1613, 2613 and so on. As a note, a new "first byte" was selected for each case file in a case in the implementation used.

Four *k* sizes were selected for testing, 512, 997, 1000 and 32749. 512 was selected as it meant a case file of 4 GB would be evenly divided into 7 812 480 samples per file with no

remainder. It is also often the same size as a sector, which means 1 byte per sector would be examined. 1000 was chosen as a "round" decimal number and 997 was chosen as it is the closest prime number to 1000. Finally 32749 was chosen as the closest prime number below 32768 which may be the upper limit of the rand() function depending on the implementation of the C library (GNU C). The prime numbers 997 and 32749 were selected after the observation that the results of testing SCSI_03 produced different patterns than the original Spike Effect program produced. The margin of error when using k of 32749 on a 4 GB file is only ±0.28% as opposed to a k of 1000 on a same sized file which is ±0.05%.

Appendix J presents six figures (Figures 13-18) for the comparative processing times using a systematic sampling approach. Figures 13-18 compare four runs of the batch script that would run the Spike Effect variant with the appropriate k value (512, 997, 1000, 32749 in that order) across six different hard drives sizes, IDE_39, IDE_44, IDE_46, IDE_71, IDE_72 and SCSI_03. The batch file processed the k = 512 samples of each of the 6 drives first, k = 997 second, k = 1000 third and the k = 32749 last. Each sampling was run independently rather than concurrently to ensure that attempting to read and process different files at the same time would not cause conflicts.

One *FTK* processing time per drive was added to be consistent for comparative purposes. In the case of the *FTK* processing time for IDE_71, the processing time for IDE_72 was selected due to the drives being the same size, and due to IDE_71 taking over 4 hours for *FTK* to process whereas the *FTK* processing time for IDE_72 was 2081 seconds, or 34 minutes and 41 seconds.

In general there are two main observations from the tables, the first is that with k = 512 the processing times were somewhat inconsistent, whereas k = 32749 were highly consistent across all drives. The second observation is that with k = 32749 the processing times were always consistently below the equivalent *FTK* processing time. It was considered (however how unlikely) that the k = 512 processing times may have been affected by being executed first in the batch processing. A second set of experiments were conducted with a batch script that only did k = 512 and k = 32749 processing, but with k = 32749 executed first on all 6 drives. The results of those two experiments can be seen in Appendix J, Figure 19. Again, the k = 32749 processing times were consistent, both through the two runs of the experiment but also when compared with the other k = 32749 values. In other words, consulting Figure 19 shows that for IDE_39, the processing time was either 262 seconds or 263 seconds, and

Figure 18 shows that both times were 263 seconds. Despite the processing order being changed for k = 512 the processing times were still inconsistent both when considering Figure 13 and Figure 18 for example.

After examining the processing times, the next important consideration was "Do the results of sampling match the results of the entire population?"

The counts and percentages for each of the six drives examined were compared with the original Spike Effect output when considering "expected values". In the case of IDE_44 it was expected that each sample would have exactly one spike representing 100% of the data, and this was used as a control for testing the samplings. If the results of sampling IDE_44 did not match one spike at 100% then there was a very clear problem with the sampling (or the program implementation). In testing all 4 k values across all 4 runs of the experiment, the results did match one spike of 100% for IDE_44.

Likewise drives IDE_46 and IDE_71 were expected to have 256 spikes each and have 1.36% and 91.4% of space accounted for by the first three spikes respectively. Across the sampling, both drives always had 256 spikes each, and the space accounted for values did not present any anomalies.

The following table (Table 7) shows the spikes found in IDE_39. Due to the random start point of each run it means different runs of the same k value produced different spike counts. However the space after accounting for the three largest spikes was 99.9999% across all the samples. As 0.0001% represents 1 megabyte on a 10 gigabyte drive this was considered initially to be close enough to the 1 megabyte threshold considered for the original Spike Effect program due to rounding. Referring back to the original Spike Effect output for IDE_39 it shows that after accounting for the three largest spikes that the remaining larger spikes had counts as high as 50 and typically counts ranged from 10-30.

Table 7 Counts of Spikes found when sampling IDE_39

Run	k =512	k = 997	k = 1000	k = 32749
1	4	6	7	3
2	6	4	4	3
3	6	6	7	3
4	5	4	7	3

IDE_72 had 137 spikes in the original Spike Effect program output, however the first spike accounted for 99.9853% of the space, (leaving approximately 11 megabytes on an 80 gigabyte drive) and the second spike covered functionally the remainder of the data. The remaining 135 spikes had between 1 and 10 counts in them each. As Table 8 shows, of the 16 samples taken of IDE_72, seven found 3 spikes and nine found 2 spikes.

Table 8 Counts of Spikes found when sampling IDE_72

Run	k =512	k = 997	k = 1000	k = 32749
1	2	3	2	2
2	3	2	2	2
3	3	3	2	2
4	3	3	3	2

The space accounted after considering 3 spikes was still 100% which means the drives were considered to be wiped as per the original Spike Effect classification.

The effect of sampling became more apparent when considering SCSI_03. The results were also considered interesting as there was a pattern on the drive. SCSI_03 had 114 spikes when analysed with the original Spike Effect program, however most of those were found to be characters used in the Master Boot Record which also explained that a pattern had been used to wipe the drive. As stated above the pattern was of 6 tall spikes of approximately 9% each, and 10 smaller spikes of 4.5% each, and there was an additional spike with a count of 8000. Therefore there were 16 main spikes, 1 much smaller spike that potentially may be found, and as the majority of the other bytes were part of the MBR then only one spike would be likely found there as the MBR is typically 512 bytes. Table 9 shows the count of spikes when sampling SCSI_03.

Table 9 Counts of Spikes found when sampling SCSI_03

Run	k =512	k = 997	k = 1000	k = 32749
1	8	17	8	16
2	12	17	6	16
3	16	17	8	18
4	8	17	8	17

As the results show however when using an even numbered *k* value typically only half the pattern was detected. This is one of the major reasons prime numbers such as 997 and 32749 were selected for examining the data as it means patterns will be potentially detected.

In all the sampling experiments, the samples detected the same overall classification that the original Spike Effect program detected (based on space accounted for by the three largest spikes), with SCSI_03 producing markedly different percentages on a drive that had been wiped with a pattern. However, it still would have classified SCSI_03 as being a special case as the maximum accounted for by the three largest spikes (18.18% each) was only 54.54%.

When considering spike counts, k = 32749 was faster in these experiments than the *FTK* processing time, and in the cases of IDE_44, IDE_46, and IDE_71 it found the correct number of spikes each time (1,256,256 respectively). In cases IDE_39 and IDE_72 it found the significant number of spikes (3 and 2 respectively) each time and it was only with SCSI_03 that the results were inconsistent and k = 997 found the significant number of spikes (17) consistently. Overall, the k = 32749 currently appears to be an acceptable sampling approach as it accurately detects the original Spike Effect classification, and is consistently faster than *FTK* when comparing drives of the same size. The margin of error is also well below ±1% which is considered appropriate for this research.

4.8.4.2 Simple Random sampling

The second variant was "simple random sampling". In this case the size of the case file was calculated, and then divided by k to calculate the number of bytes to be analysed. Two approaches were then trialed. The first approach was to repeatedly randomly pick a byte from the file, read it in and analyse it. The concern with this approach was the potential for taking longer due to moving the file pointer back and forward around the file, and the possibility of analysing the same byte more than once. The second approach was to create a list of random unique bytes, sort that list in ascending order and then analyse the bytes in the list. The concern with this approach was the possible time taken to generate and sort a large list (approximately 4 million with k = 1000) of random numbers and guarantee the uniqueness of each number. As mentioned, some rand() function implementations cannot pick random numbers greater than 4 billion which would prove troublesome for SCSI_03 which was one 18 GB file. The solution therefore was to use a random number generator that could handle much larger numbers. Both approaches for "simple random sampling" produced processing times that took well over the 2 minutes 43 seconds per 4 gigabyte that the regular Spike

Effect program took and were rejected at this time due to the concerns raised. A better implementation of the Spike Effect program may mean that a simple random sampling variant can also be implemented successfully.

4.8.4.3 Other potential approaches instead of sampling

Other approaches that maintain the sampling of all bytes in a case could be to change the language used, restructuring the code to be more efficient or to read and process a sector (512 bytes in this example) at a time and compare the sector with known or pre-calculated sectors. In other words, if the sector that was read contained all "Null" bytes then 0 is incremented by 512 once, rather than individually 512 times. Those considerations were not initially addressed in the design of the program as it was intended to be a byte histogram generator that could process multiple evidence files accurately and be scripted to process multiple cases typically overnight.

4.8.5 Future Work for the Spike Effect

As outlined above, the main areas of future work for the Spike Effect program and analysis technique include improving the original program efficiency to see if it can be as fast as or faster than *FTK*. That may be possibly achieved by using a different programming language or implementing other comparison techniques. The use of sampling approaches can be tested on all the existing case files and determining if sampling produces incorrect results. Other techniques would be applying machine learning techniques for the determination of the drive classification. Finally there are a number of issues that can improve automation, and the addition of a graphic user interface may be practical.

4.9 Analysing non-wiped drives

With regards to analysing the non-wiped drives there are two considerations, the choice of analysis tool, and which data to analyse. Furthermore when considering which data to analyse there is also the order in which those data are analysed.

4.9.1 Choice of analysis tool

As prior research shows there are a number of tools that can be used for analysing captured evidence files including *Encase* and *FTK* (both available to the author). The *Forensics Tool Kit* (*FTK*) was the software tool of choice for the analysis of the evidence files for this research partly due to the ease of use and information displayed (see Figure 9 on the next page).

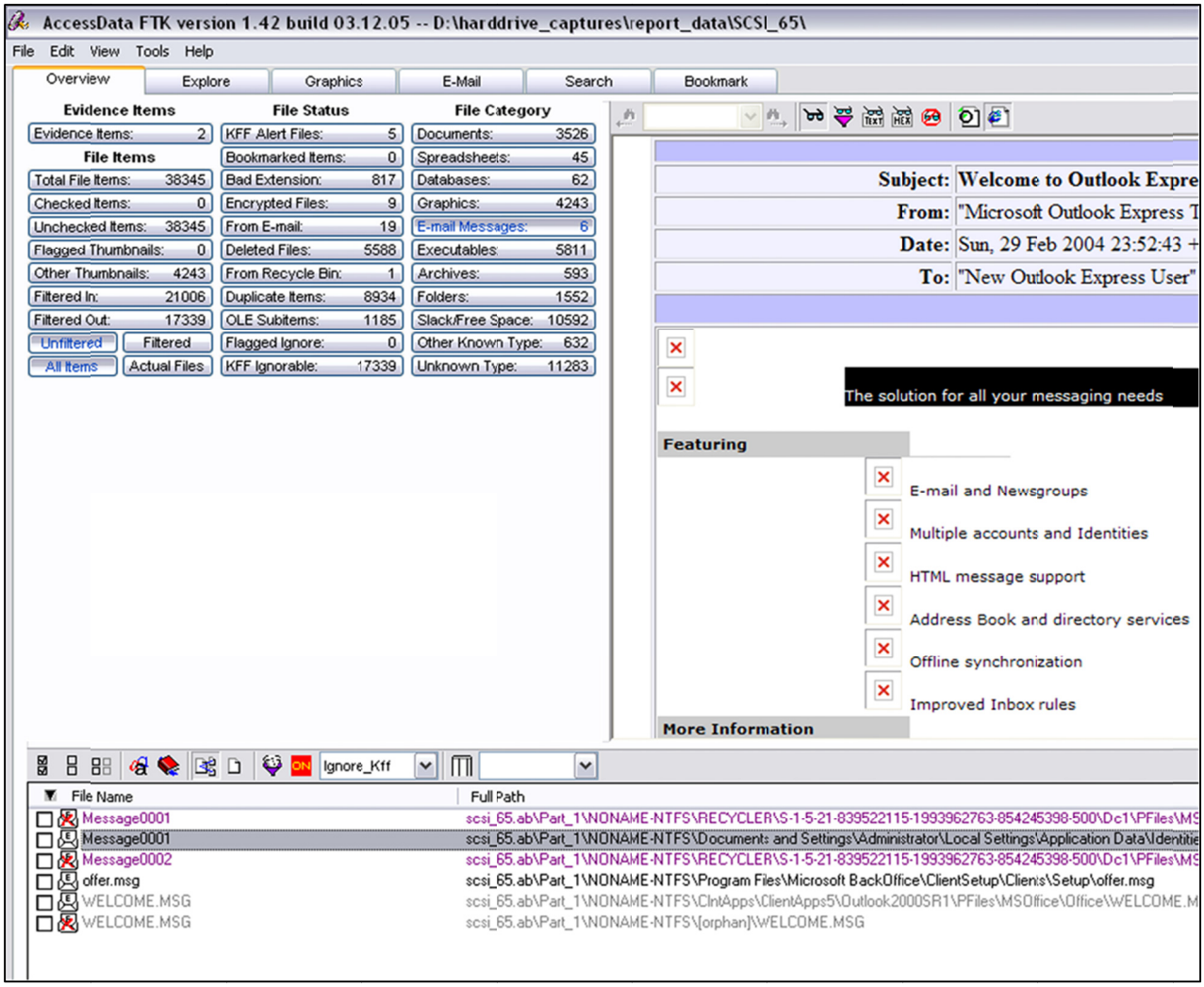


Figure 9 ***Forensics Tool Kit*** **(*FTK*) screenshot**

One of the advantages of *FTK* is that it classifies the data found into the file categories it was found in. From Figure 9 it is clear there are six email messages found, and the locations and state of the emails are displayed in the bottom half of the screenshot (three have been deleted and marked with a red x) and there is a reading pane shows that the actual email message on the right hand side.

Additionally, the software allows the use of Known File Filtering (KFF) which as Figure 9 shows there were 17339 files known to *FTK* as special files and can safely be ignored. The author did not have access to the law enforcement special filtering libraries to examine the drives for objectionable material and as image files were often not examined the decision was made not to include a count of "illicit material". Finally another feature of *FTK* is the data carving (aka file carving) feature which is easy to use and displays the carved files in an easy to understand manner.

4.9.2 Determining which data to analyse

When considering which order to examine files in, the following quote (circa 1600) which is often attributed to Cardinal Richelieu was considered:

> "If you give me six lines written by the hand of the most honest of men, I will find something in them which will hang him" (WikiQuotes, 2011)

The methodology for finding identifying information for this research was therefore to start by checking emails, then spreadsheets, and then documents. Graphics files were examined if there appeared to be a conflict or lack of clarity over the identifying information, often where it appeared to be one or more companies or one or more individuals mentioned, or in some situations where other contextual clues (such as statements in emails, logos embedded in word documents, or folder names like "Smith's holiday pix") suggested there may be identifying information that had not been determined beforehand.

This decision to generally ignore image files was made in part from the methodology considerations presented by El Emam et al. (2007) with regards to illicit images and also due to the fact that in the majority of cases in this research, there were few emails, many documents and a very large number of graphics. In many cases the images found could not be classified as identifying as the following example shows.

Figure 10 is a hypothetical example of a flag JPEG that had been deleted, and then reconstructed using data carving.

Figure 10 Data carved Flag

The data carving was not particularly successful and the image is rendered with the majority blue colour for most of the image. In other data carving attempts, the image might have large blocks of red, blue and white where the stars should be. From this data carved example it is possible to determine that the image is likely to be a flag, but it would not be possible to

determine if it was the New Zealand flag, the Australian flag, a different flag such as the Fijian flag, or it may be part of a poster or other image file entirely.

One of the reasons the flag (in Figure 10) is not identifiable (when compared with Figure 11 and Figure 12, this page and the next page respectively) is described by two of the potential limitations of file carving by Garfinkel (2007, S2):

> First and most important, these programs can only carve data files that are contiguous – that is, they can only create new carved files by extracting sequential ranges of bytes from the original image file. Second, carvers do not perform extensive validation on the files that they carve and, as a result, present the examiner with many false positives – files that the carver presents as intact data files, but which in fact contain invalid data and cannot be displayed.

So for Figure 10, the data carving process would have been able to determine the overall size of the JPEG and could find the first quarter of the data and used the last colour found to fill the rest. Garfinkel's description of the second limitation was another reason why this thesis did not use the "file count" approach for reporting that Medlin and Cazier used. Often there would be many file fragments that *FTK*'s data carving process could find and classify as documents but when examined those files did not contain viewable or meaningful data.

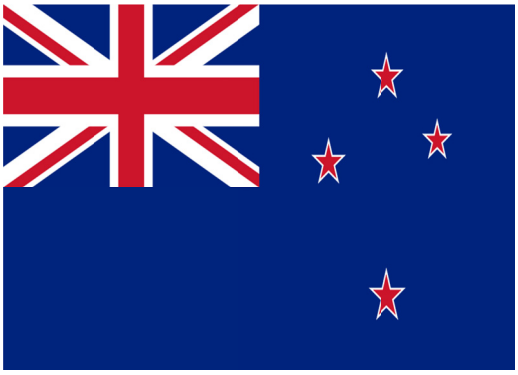


Figure 11 Flag of New Zealand

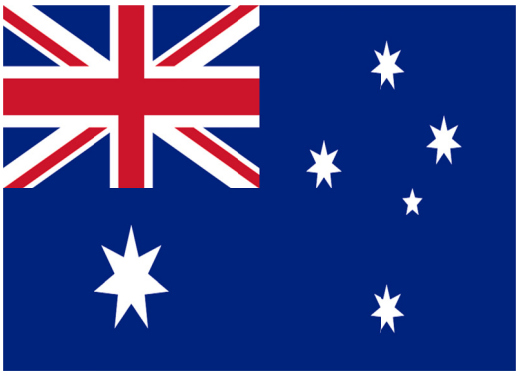


Figure 12 Flag of Australia

For example, finding the flag shown in Figure 11 may suggest that the user is a New Zealander, while the flag shown in Figure 12 may suggest the user is an Australian. Alternatively it may suggest the user has gone to a New Zealand or Australian themed website and the file is in their cache, or that they saved the file for an assignment, report or presentation they were working on.

This is one example of why image files often do not provide identifying information as they often have subjective meanings, or they lack a clear connection or context between the user(s) and the information in the image. That also assumes the image was found in a complete state.

4.10 Reporting the data found

There are two main reporting styles used in the prior work in data remanence. The first style is to count each instance of specific file types such as "word documents", "emails" and the like and report those values (Garfinkel & Shelat, 2003); (Medlin & Cazier, 2011). From Figure 9 (on page 94) getting approximations of those values would be easy using *FTK* but it would also be unclear just from that raw data the nature of the identifying information. In a number of cases with this research, some hard drives had over 400 000 files which would skew the results when compared with hard drives of smaller capacity. Additionally as described by Garfinkel (Garfinkel, 2007), the file counts could be distorted with false positives.

The second style used by the consortium (Jones, et al., 2010) attempt to answer two questions with their reporting style; "How many drives are wiped?" and "How many drives had information that could identify companies and/or individuals from the remnant data found?"

First the total number of drives is displayed, then the number of unreadable and wiped drives. This then leaves a total of drives that have readable data on them. The challenge with understanding the consortium style is that they subtract the unreadable drives from the sample size, and then the wiped drives when reporting which mean the percentages do not always add up to 100%. Secondly the consortium then also present the counts of company identifying and individual information found (in the same table) but do not have a count of drives found that had both types of information. Table 10 provides an example of the reporting style from the consortium 2009 results (Jones, et al., 2010).

Table 10 Example of Consortium reporting

Country	Total	Unreadable	Wiped	Remaining	Company Identifying	Individual Identifying	Formatted	Illicit
UK	174	60 34%	32 28%	82	28 34%	36 45%	18 22%	2 2%

For example "174 drives in total, 60 drives are unreadable (34%). 32 drives have been analysed as wiped (32/ (174 – 60)) remaining readable drives or 28%". Therefore there are 82 drives remaining that are not wiped.

The consortium then report how many drives had had attempts at removing/hiding the data such as formatting, deleting, re-partitioning, or operating systems re-installed (which they refer to as "Formatted" but this thesis refers to as "Reformatted") and how many drives had illicit material on them.

Continuing the above example there were 28 drives of the 82 remaining drives (34%) with company identifying information, 36 of 82 remaining drives (45%) with individual identifying information. Looking across the table however, the total percentages add up to 165% which as mentioned can introduce some confusion if the reporting method is not fully understood.

It is possible for a hard drive to be any combination of those four values (Company Identifying, Individual Identifying, Formatted and Illicit), so one drive might be individual identifying only, whereas another drive may be company identifying and also have been

formatted. One of the challenges with the way the consortium present identifying information is that it is not clear the total number of drives that have identifying information on them.

This thesis considered how to present the information found such that it could be comparable with the consortium where possible, be easy to understand and present the information that the consortium do not present. The first change was to split the results into two tables instead of one to represent the two main questions this author was interested in. Table 11 is a template for the first table and first question "How many drives were wiped?" Unlike the consortium tables, Table 11 is designed so that the three rightmost values will add up to the "Total" value and percentages will add up to 100%.

Table 12 (which appears on the next page) addresses the question of "How many drives had information that could identify companies and/or individuals from the remnant data found?" which is often shortened to "How many drives had identifying information on them?" The main changes with this table are a field for "Both" to represent drives that had both "Company Identifying" and "Individual Identifying" information on them, "CI" represents drives that only had "Company Identifying" information and "II" represents drives that only had "Individual identifying" information on them. Those fields were then divided into "Clearly Viewable" and "Reformatted" respectively to make it much clearer if there appeared to have been efforts taken to remove the data. Although the fields in "With Identifying Information" in Table 13 are in alphabetical order, they are also considered by this author to be in order of higher risk to lower risk. That is, a hard drive where the data are "Clearly Viewable" and has "Both" company and individual identifying information is potentially more of a risk than a drive that has at least been "Reformatted" and only has "Individual Identifying" information on it. Naturally the type and nature of the data would actually determine the risk. This research does not include a count of "illicit material found" therefore there was no consideration for that in the new tables.

Table 11 Template table for answering "How many drives were wiped?"

Name	Total	Unreadable	Readable	
			Wiped	With Data
Total				

Table 12 Template table for answering "How many drives had identifying information?"

Name	Total	Without Identifying Information	With Identifying Information					
			Clearly Viewable			Reformatted		
			Both	CI	II	Both	CI	II
Total								

Roberts and Wolfe (2011) presented the results in the consortium style for compatibility with the consortium results however the templates above are used in the next chapter to present the data and results for this research.

4.11 Methodology Summary

The methodology for this experiment was designed by examining prior data remanence experiments. Six objective comparable elements of the prior methodologies were identified and four further elements such as sample size, reporting, identifying information and research questions were discussed.

A flow chart for conducting the experiments was developed and presented in Figure 2 (on page 65).

A piece of software (the Spike Effect) was created as part of this research to attempt to automate part of the data processing. It was successful in that it was able to determine which hard drives had been wiped and therefore did not need further examination and this saved a lot of time overall. Variants of the Spike Effect that used sampling were also tested as it was noted the original Spike Effect was not as fast as *FTK* (the analysis program used) in certain cases.

A new reporting style was developed to clarify information that was not clear from previous reporting styles.

Chapter 5: Analysis of the Data

5.1 Introduction

This chapter considers the 100 hard drives examined in the experiments, the states of the hard drives (unreadable, wiped, readable) and the types of data found (company identifying, individual identifying and drives that had both types) to answer the primary and secondary research objectives. Appendix C contains the descriptions of the hard drives and data found on them.

The analysis will then be used to answer the primary and second research objectives. The question of *Tableau*'s "Source drive may be blank" will also be considered in this section.

Other researchers often do not state the interfaces of the hard drives they examine or apply considerations if the interface type played any significant role in their results and this chapter will start by examining the interface type to determine if this had a relevant role in the results.

5.2 Sources of the drives

The drives from the private companies were sourced "blind" to the author for analysis by the author's supervisor who in turn had made arrangements with Company A and Company B. Both companies agreed to loan "unknown state" drives on the basis that any hard drive supplied would be forensically wiped before being returned and that the author would pay all shipping costs. A secondary requirement was that any drive that was unreadable by the system used for the experiments would be marked as such when returned. A total of 50 hard drives were supplied by Company A in three shipments with the understanding that they would supply as many drives as they were able to.

Much like the arrangement with Company A it was understood Company B would supply as many drives as they were able to. Company B supplied a total of 19 drives in one shipment. Initially it had been understood there would be 20 drives and it was not satisfactorily determined why the shipment was short one drive. Company B did not send any further shipments. This author is grateful for the support with this research by Company A and Company B.

The auction site drives were all purchased from Trade Me. The drives were selected from lots on Trade Me that were "unknown state" drives and the author and his supervisor reviewed

the lots before purchasing them. This was due to in part to budgetary constraints but also due to the ambiguous wording of some lots either the interface type of the drive, or wording used to describe the potential state of the data on the drives. As per recommendations from other researchers (Jones, et al., 2010, pg. 93) third parties were asked to place bids or use the "buy now" options on the Trade Me auction lots to make it less obvious that data remanence research was being conducted. The author is grateful to those third parties for donating their time for this project.

5.3 Observations and changes from the methodology

There were no major changes from the methodology. The only two issues of note were that Company A provided a full range of hard drives including SCSI and SCSI-SCA interface drives which had not initially been expected. This was a potential challenge as the *Tableau* did not have an interface for those drives at the time of experiments. This was resolved by adding a SCSI card and cabling to a normal desktop computer and by using the *FTK Imager* software.

The second issue is that Company B provided less hard drives than initially expected. As the consortium results have included counts of sources that were under 30 (North America, 24 drives in 2006, Table 28, Germany 28 drives in 2008, Table 30, France, 17 drives in 2009 Table 31, as shown in Appendix I), it made sense to include the results from Company B with the caveat that it was less than had been planned for. Conversely, the total number of drives from Trade Me vendors was 31 when only 29 had been ordered.

5.4 Brief Results

Tables 13 and 14 (below and on the next page respectively) show the 100 drives analysed for this thesis. Table 13 considers the initial data remanence question of "How many drives in a sample were wiped?" and Table 14 considers "How many drives have identifying information on them?"

Table 13 Total of all New Zealand Hard Drives analysed

Name	Total	Unreadable	Readable	
			Wiped	With Data
Total	100 (100%)	21 (21%)	34 (34%)	45 (45%)

Table 14 Breakdown of Identifying Information as found on New Zealand Hard drives

Name	Total	Without Identifying Information	With Identifying Information					
			Clearly Viewable			Reformatted		
			Both	CI	II	Both	CI	II
Total	100	76	10	1	2	4	4	3

As Table 14 shows, 34 drives of 100 were wiped which meant it was not possible to find any remaining data at all. As Table 15 shows, 10 drives of the entire sample of 100 had both company identifying and individual identifying information on them and the data were clearly viewable on those drives, no special tools would have been needed to find that data for a typical user. A secondary observation is that coincidentally the total number of drives that had individual identifying information is 19 (Clearly Viewable Both + Clearly Viewable II + Reformatted Both and Reformatted II) which is the same as the total number of drives with company identifying information (Clearly Viewable Both + Clearly Viewable CI + Reformatted Both + Reformatted CI).

5.5 Hard drives of interest

Full descriptions and notes can be found in Appendix C of the results of the hard drives being analysed. A number of drives from Company A had been wiped with patterns, similar to SCSI_03 which showed the benefits of using the Spike Effect program. Two drives from Company B were of interest those being IDE_59 and IDE_69. From the Trade Me auctions there were also two drives of notes: IDE_78 and IDE_86.

IDE_59 was a 40G hard drive, and was a drive that *Tableau* stated "Source drive may be blank". The "Spike Effect" for this drive showed all 256 characters represented and approximately 87% for the first spike, and the other 2 spikes being 0.5% and 0.25% respectively. This meant there was potentially 5 gigabytes of data left on the drive. Analysing with *FTK* allowed for a "disk view" which showed a list of user names in the format of full first name and full last name i.e. John_Doe. When emails were examined there were a number of business quotes to customers. Examining one of the user directories however found one of the most interesting pieces of data in this research.

The user had signed up for a virtual sports betting site which is legal within New Zealand as no real money is used. The web form had asked for (and stored) the following values "User

name", "First name" "Last name", "Street address". "Email address" "Phone number" "Cell phone number" and "Date of birth". There was a "password" field however it did not store the given password and no attempts were made to recover the password.

The user had put in an email address in the format of John Doe at a company name which was consistent with the company named elsewhere on the hard drive. The user's name was also consistent with data found on the hard drive. The other details were all consistent with street addresses and telephone area codes in use in New Zealand. Not only was the data of concern given a credible date of birth was provided, other information found on the hard drive [email dates, web page information] showed the data strongly appears to date from 2009. Overall this was one of the most important drives found in this research as it had many multiple clear examples of identifying information (as there was additional data not found in public directories such as a telephone directory) and due to the timeliness of it, also shows the risks from data remanence as described in Section 2.6 on page 32.

IDE_69 was also clearly visible and had OS and files on it. After examining the files there was nothing identifying so data carving was used. The files returned contained psychometric tests for three employees (names and then relative attributes in a variety of situations such if the person being tested works well in group situations, has leadership potential, is honest and such like). Of note with the psychometric test however was the word "Confidential" at the bottom of each page, and references to the 1993 Privacy Act such that the employers needed to concern themselves with the proper storage and deletion of the tests. This is another example of the risks of data remanence and the need for companies to ensure they dispose of their data properly.

IDE_78 had a number of issues when *FTK* attempted to process the forensic copy case files which meant it could not be fully processed. The processing had to be manually stopped once *FTK* had processed approximately 280 000 files (typically 22 hours of processing time). It should be noted *FTK* had successfully processed drives that had over 400 000 files as part of this research so the volume of files alone was not the issue. Multiple attempts at the processing had been done before the decision was made to stop the processing at the 280 000 point. No adequate reason could be determined for why the processing would not complete for this drive. It appears to have been a server machine and whilst it definitely had company identifying data, it was not possible to find individual identifying information in the 280 000 files processed.

IDE_86 was a reformatted hard drive. Data carving found a number of web pages that had interesting individual information. The main web form that the owner of the hard drive interacted with was an Australian lottery syndicate which initially appeared to let users purchase tickets for a number of Australian lotteries. On subsequent pages the user was encouraged to sign up friends and family, and the initial user would then get discounts and/or benefits from those signing up. This initially appeared to either be multi-level marketing which is legal in Australia and New Zealand (under certain conditions) or a pyramid scheme (which is illegal in both countries). The web pages for that site also showed that the user had deposited money into their syndicate account a number of times, and the credit card number was partially obscured (it was difficult to tell if this was a design feature or if it was how the html rendered the credit card details onto the screen). Because it is unclear if it would have been illegal in New Zealand and because the information was dated to 2003 the decision was made not to contact the New Zealand Police. This drive does however highlight the potential risk of blackmail to the former owner if the syndicate activity had been illegal.

5.6. Analysis of results by interface type

5.6.1 Introduction

Because other data remanence researchers generally do not publish the types of the interfaces in their research, it is not initially clear what factor if any the hard drive interface would play in the overall data remanence baseline. An exception to this was Valli and Woodward (2008) which was targeted at company identifying information and involved purchasing SCSI drives. Section 2.2.2.3 (on page 12) outlines that the hard drive interface such as SCSI, SATA and IDE may not be a good indication of the role of the hard drive (either as a server drive or a home user drive). As examples SCSI_35 was a clearly viewable drive that had a server operating system on it, and IDE_80 (which was reformatted) strongly appeared to be a web server due to the name and type of html files, associated image files and other relevant files. IDE_81 was definitely a home user drive as it had computer games, the Windows 98 operating system and no company identifying information on it.

Both Company A and Company B sourced drives from servers, business work stations and home user hard drives and sent a random blind selection for this research therefore it was not possible to determine which drives had come from servers and which had not from those vendors.

It was not always possible to determine the age of the hard drives (such information either not printed on the drives or not clearly visible, or not determinable from the data on the drive itself) which means it was not always possible to determine if the drive had a SECURE ERASE function built-in. The inability to accurately date the hard drives also meant it was not necessarily possible to determine which drives were less likely to be readable due to age related issues. A number of comparisons by interface however can be drawn to determine if the interface made a difference to the experiments. Although there were only three SATA drives they have been included for comparisons with future research. Table 15 summarises both the types of interface and their sources which will be used in the following analysis.

Table 15 Summary of Interface by Source

Source	IDE	SATA	SCSI	Total
Company A	12	3	35	50
Company B	15	0	4	19
Trade Me	31	0	0	31
Total	58	3	39	100

5.6.2 Unreadable drives by Interface:

As stated by Anderson et al. (2003) SCSI drives are expected to be used 24 hours a day 7 days a week as server drives whereas IDE drives are more likely to be home user hard drives that were used 8 hours a day 7 days a week or less.

Table 16 shows the comparison between interface types where the drive was unreadable.

Table 16 Summary of Unreadable drives by Interface

Interface type	Count of unreadable drives	Total of that interface type	Percentage of unreadable drives by interface type
IDE	12	58	21%
SATA	1	3	33%
SCSI	8	39	21%
TOTAL	21	100	21%

When comparing unreadable drives with other drives of the same interface type, SCSI and IDE were tied at 21% meaning roughly 1 in 5 drives per interface type were unreadable. When considering the percentage contribution of each interface to the total of unreadable drives, IDE drives were 1.5 times more likely to be unreadable than SCSI drives however.

5.6.3 Wiped Drive by Interface

The next comparison that can be made is considering the interface of a wiped drive. Table 17 (on the next page) shows the comparison between interface type where the drive was wiped. The values in table have been derived by removing the counts of unreadable drives from each interface count.

Table 17 Comparison of Interface type on wiped drives

Interface type	Count of wiped drives	Total of that interface type	Percentage of wiped drives by interface type
IDE	18	58	31%
SATA	1	3	33%
SCSI	15	39	38%
TOTAL	34	100	43%

When comparing like with like, just under 50% of the SCSI drives were wiped and just under 40% of IDE drives were wiped. When considering the percentage contribution of each interface to the unreadable drives the results show that IDE drives were only 20% more likely to be wiped than SCSI drives.

SATA drives natively support a SECURE ERASE function which means it is potentially easier to remove the data securely from the drives. However as of 2006 SCSI did not support this feature, and IDE drives may or may not have the SECURE ERASE feature depending on when they were manufactured. Due to not being able to consistently determine the age of the hard drive for IDE drives it was not clear if the drive would have supported SECURE ERASE.

5.6.4 Interface comparison by Identifying information

The final interface comparison is for identifying information. The data has been displayed comparing the interfaces over all three types (both, company identifying, individual identifying) and shown in Table 18.

Table 18 Comparison of Interface and Identifying Information

Interface Type	Total	Readable drives with remaining data	With Identifying Information		
			Both	CI	II
IDE	58	28	13	4	4
SATA	3	1	0	0	0
SCSI	39	16	1	1	1
Total	100	45	14	5	5

As stated in Section 4.3.2.2 on page 60, a margin of error of ±20% or greater was considered a threshold for concern. For the observations regarding IDE, the margin of error is less than ±15% but larger than ±10% and for SCSI it is approximately ±20%. SATA as stated was included for completeness and further research.

From Table 18 it is clear that IDE drives had the most identifying information on them. IDE drives comprised 47% (21 identifying information drives out of 45 remaining drives) of the entire set of readable drives. 75% of all the readable IDE drives with remaining data (21 of 28 drives) had identifying information on them. A secondary observation is of the 21 IDE drives that had identifying information on them, 17 had Company Identifying information on them (Total of "Both" and "CI" in Table 18 on the previous page). It should be noted that it is not possible to determine how many of those were server drives and how many were desktop drives used by employees at the workplace.

The sample size of SCSI drives with readable remaining data at 16 would have a margin of error of approximately 20% (±3 drives in this case) therefore it is difficult to make adequate evaluations for SCSI drives. In this particular case the range would have been 3 drives ±3 drives or 0-6 drives. A recommendation therefore would be to examine more SCSI drives only, and then more SCSI drives in the overall research pool. As such the author and his supervisor have been invited to participate in a consortium based research project which will be targeted research as it focuses on company profiling by sourcing SCSI drives in 2012 from auction sites. That research project is contingent on funding and availability of time however would likely provide an answer to the above question of SCSI drives when considering identifying information as many more SCSI drives will be examined and have a more suitable sample size.

5.6.5 Summary of comparisons by interface

Overall, the interface of the drive was only relevant when considering the issue of identifying information for this research. In the case of the SCSI drives identifying information had either been adequately removed, or in some instances may never have existed on the drives in the first place. Conversely most of the readable IDE drives had identifying information on them. This suggests that for researchers conducting targeted research in New Zealand, buying IDE drives is more likely to be successful, and secondly that whilst all users should be encouraged to adequately remove their data before selling, that education should be focused on IDE drive sellers. The caveat of course remains that for each interface type a larger sample

size is recommended as the samples for this research were a pilot study and had a margin of error of at least ±10%.

5.7 Analysis by source

5.7.1 Introduction

This section considers if any particular source of hard drives for this research had a greater impact on the results and if so the relevance of that impact. As there are five main category of results ("unreadable", "wiped", "company identifying", "individual identifying" and "total identifying") it is possible that different sources influence some categories but not others.

One factor for special consideration is "the reliability of the source" because Trade Me allows traders to rate the trades between the buyer and seller, and this feedback is immediate and available for users of Trade Me to read. The feedback from other traders with Company A and Company B is less immediate and is not available online. Table 19 presents the summary of drives based on their source and the state of data found on them.

Table 19 Summary of drive source and drive state

Name	Total	Unreadable	Readable	
			Wiped	With Data
Company A	50	9 (18%)	26 (52%)	15 (30%)
Company B	19	9 (47%)	0 (0%)	10 (53%)
Trade Me	31	3 (10%)	8 (26%)	20 (64%)
Total	100 (100%)	21 (21%)	34 (34%)	45 (45%)

The margins of error are approximately ±10% for Company A (50 drives out of a sample of 100), ±20% for Company B, ±15% for Trade Me.

5.7.2 Unreadable drives by source

Considering "Unreadable" drives, whilst Company A and Company B tied for raw count of unreadable drives, in percentage terms almost half of Company B drives were unreadable whereas less than 1 in 5 Company A drives were unreadable. Furthermore only 1 in 10 Trade Me site drives were unreadable.

Company A's unreadable drives were predominantly SCSI drives (seven of the nine unreadable drives). Of the remaining two unreadable drives one was a SATA drive and one was an IDE drive. As Company A provided the drives in three different lots, and the second

lot had five unreadable drives then damage in transit is a valid possibility. The age of the drives based on data found on the readable drives is also another valid possibility.

The unreadable rate for Company B was almost 50%. Examining the external labeling on the hard drives and the data found on them provided insight into why this was. The drives that were readable had information that was over five years old and in some cases over 10 years old. If the drives that were unreadable were from the same source (that Company B acquired them from) or same timeframe as the readable drives then it is likely they were unreadable due to age and the related reliability problems.

The results for the Trade Me have one caveat on them. 29 drives were purchased but 31 drives were shipped. This situation had not been anticipated during the methodology design as other researchers had not reported it happening in their research.

The initial reasons for Trade Me vendors including extra drives were considered to be that the vendors were either looking to get rid of drives they knew were damaged (as the purchaser pays shipping) or they were not confident in the quality of drives that they were providing so they added drives in (much like the concept of a baker's dozen being 13 rather than 12). The other possibilities include that they had miscounted the number of drives they had for sale in the lot, had made a typographic error when creating the auction advert, the vendor was being generous in the hope of getting positive feedback for the trade or that the vendor was looking to get rid of excess or final stock.

As other researchers had not commented on it, there were no considerations in the methodology to determine the reasons for the inclusion of the extra drives. The solution was to examine the groups that had the extra drives and noting if those groups had unreadable drives or not.

The two auction lots that had extra drives were examined to determine if the first two reasons were more likely (Vendor purposefully added an unreadable drive or vendor added a drive because they were not confident all drives were readable). One of the extra drives was in the third group of drives shipped. In that group there were two unreadable drives so the first two reasons are potentially valid. One of the extra drives was in a fourth group that was sent, and those six drives were all readable and all wiped this suggests the first two reasons are unlikely to be valid.

As only three drives (10%) of the sample of Trade Me drives were unreadable (and one was possibly thought to be unreadable by the vendor when it was sent) this means that the Trade Me traders were very reliable in sending working drives. One of the reasons as outlined above is that traders can receive feedback quickly that positively or negatively affects a feedback score they have and it is public information therefore providing good reliable service is beneficial to the trader. Future work from this research would be to conduct a longitudinal study primarily from auction sites for both "known" and "unknown" state drives to determine if the low unreadable rate is consistent and if there are other factors that affect that low value.

5.7.2 Wiped drives by source

The next consideration is wiped drives relevant to the source of the drives.

Company A had 10 IDE, 1 SATA and 16 SCSI drives wiped for a total of 63% of the readable drives. From Company A this means it was the norm for drives to have been wiped rather than not wiped. This raises the possibilities that Company A had been sourcing drives from companies and individuals that were aware of the data remanence problem and had taken actions to protect their data, that Company A recommends their clients wipe the drives before selling them to Company A or that Company A provides the service of wiping drives. As part of the original discussions (as reported to the author by his supervisor) with Company A they had stated they neither wipe drives nor advise their clients to wipe the drives before selling them. This is reinforced by the agreement with Company A that all drives sourced for the research from them would be wiped at the end of the research before being returned to Company A.

Company A's clients are nationwide and do not share characteristics that suggest they would be more aware of the data remanence issue than other users. In other words Company A does not exclusively source drives or computer systems from government agencies that should be aware of the GCSB manual and guidelines for secure data deletion. Before further conclusions could be drawn from this, the Company B hard drives also need to be considered.

Company B immediately stands out as none of the drives were wiped. Company B shares the same main attributes as Company A that is sourcing drives nationally, and do not state they wipe their clients drives after purchasing those drives, nor recommend the clients wipe their drives.

The Trade Me site had eight wiped drives, representing 29% of the readable drives from that source. As noted earlier, six of those drives came from one source and when the known state research project is conducted (Section 6.4 Future Work on page 127) it would make sense to consider that particular vendor as a "known state" vendor and purchase more drives from that source to determine if that was an anomaly for that vendor or if that vendor is doing something in particular with all drives that they sell. The auction advert for that lot did not give any indications that they did anything in particular to the drives.

It is worth noting that all auction drives can potentially come from various sources before they are sold by any auction site vendors. Therefore those traders may have received the drives already wiped, or may have wiped them themselves.

5.7.3 Identifying information by source

Table 20 presents the breakdown of identifying information found on the drives and their source.

Table 20 Summary of drive source and identifying information found

Name	Total	Without Identifying Information	With Identifying Information					
			Clearly Viewable			Reformatted		
			Both	CI	II	Both	CI	II
Company A	50	47	0	0	0	1	1	1
Company B	19	11	5	0	0	2	1	0
Trade Me	31	18	5	1	2	1	2	2
Total	100	76	10	1	2	4	4	3

The high number of wiped drives from Company A means that overall the number of totally identifying drives is low for making significant statements about the overall population let alone considering the individual sources. Conversely Company B is of note as eight of the ten readable drives did have some form of identifying information on them. The Trade Me auction lots had 13 of 20 (65%) readable drives with identifying information on them.

5.7.4 Analysis of Company A versus Company B

As the results show, both Companies A and B supplied drives that had remaining data and identifying information on them which shows that in this case that neither company consistently wiped all drives before selling them. Both companies made the claim that they do not directly tell their clients to wipe the drives before selling them. There is the consideration that much like the drives being unreadable issue most likely relating to the age

of the hard drives it is also quite possible that Company B's group of drives were sold to Company B years before Company A's drives were and therefore people selling to Company B were likely to be less aware of the problem of data remanence.

5.7.5 Summary of impact of the drive source

Overall the source of the drives had an effect on the outcome of the results. Drives sourced from Company A were "good" as almost all the drives that were readable were either wiped or had any identifying information adequately removed. Conversely Company B's drives were either unreadable or the readable drives were not wiped at all. The sample size for Company B was lower than desired as there were issues surrounding the initial shipment of 19 instead of 20 drives that meant the initial plan of acquiring 50 drives from Company B did not happen.

As the drives that were supplied by Company B met all the requirements to be included in the research (available for purchase on the second hand market, by members of the public, no statements about the nature of the data on the drives, and were hard drives rather than other types of data storage device) it was decided that the drives would be included in the final analysis.

Drives from Trade Me as an auction site created some interesting observations. Due to the overall reliability of the drives sourced from Trade Me (28 of 31) and the number of drives with identifying information on them (13 of the 28 readable drives) it suggests that Trade Me is a good source of drives for those looking to purchase drives with identifying information on them. Future research would be to firstly examine more drives from Trade Me to get a better sample size, and then to get larger samples from private companies other than Company A and Company B, and from other auction sites.

5.8 The Primary Research Objective

The primary research objective was to determine how does the baseline level of data remanence in New Zealand compare to other countries. One of the more practical ways to determine the answer of this comparison would be the compare the New Zealand results with those of the consortium 2009 (Jones, et al., 2010)results as those results were the most recent published for the consortium and cover five different countries with different population sizes, spread over geographic areas and with different attributes such as being landlocked or being islands. Another reason to consider the consortium results is that the consortium has

been conducting research consistently over a number of years whereas other researchers tended to be a "one and done" approach to data remanence research for hard drives.

There are two main ways to compare the New Zealand results with the consortium results, the first way is directly compare the New Zealand results with the consortium combined results, and the second way is to compare the results country by country. As can be seen in Appendix K however, ranking proved difficult as there were a number of outliers per category and/or the sample sizes were particularly small as well as the overall problems of sampling raised earlier. This meant the main comparisons can only be with the overall consortium results.

5.8.1 Comparison of New Zealand results and Consortium combined results

The first comparison using the New Zealand results and the consortium 2009 results from Table 1 "Amalgamated results of 2006-2009 for the consortium" (on page 44) are presented in Table 21.

Table 21 Direct comparison New Zealand vs. Consortium 2009

Name	Total	Unreadable	Readable	
			Wiped	With Data
New Zealand	100	21 (21%)	34 (34%)	45 (45%)
Consortium 2009	346	94 (27%)	89 (26%)	163 (47%)

The primary comparison for New Zealand versus the consortium was considered to be the number of wiped drives in each sample.

Using *PASW Statistics 18* generated the following crosstab (Table 22):

Table 22 Cross tabulation of wiped drives for New Zealand and the consortium

		Counts		
		Wiped	Non-Wiped	Total
Country	NZ	34	66	100
	Consortium	89	257	346
Total		123	323	446

Using a "Fisher's exact test" with the null hypothesis that the proportion of wiped drives in New Zealand is the same as the consortium produced a two tailed $p = 0.127$.

Typically $p < 0.05$ means reject the null hypothesis. Therefore $p = 0.127$ means do not reject the null hypothesis. When using "Fisher's exact test" for the null hypothesis that unreadable drives in New Zealand is the same as the consortium produced a two tailed $p = 0.244$ and for completeness the readable drives comparison produced a two tailed $p = 0.734$. As all three null hypotheses can be accepted then it is considered that the overall hypothesis that the NZ drives are comparable with the consortium drives is accepted.

Comparing identifying information rates requires considering the number of identifying drives within the sample of readable drives for both New Zealand and the consortium. The New Zealand counts were derived from Table 21 on the previous page. A similar cross tabulation is shown in Table 23.

Table 23 Cross tabulation for company identifying information comparison

		Counts		
		Company Identifying	Not Company Identifying	Total
Country	NZ	19	26	45
	Consortium	65	98	163
Total		84	124	208

Using a null hypothesis that the proportion of company identifying drives in New Zealand is the same as the consortium produced a $p = 0.877$ so the null hypothesis is not rejected. Likewise using a null hypothesis that the proportion of company identifying drives in New Zealand is the same as the consortium produced a $p = 0.876$ value so it is also not rejected.

The important consideration regarding identifying information is that while a drive must be either "unreadable", "wiped" or "with data" a drive could have both company identifying and individual identifying information. Furthermore the New Zealand results showed that drives with identifying information tended to have both types of information rather than just one type. The same observation cannot be made for the consortium as they do not list a count for drives that have both types of identifying information.

5.8.2 New Zealand results compared versus individual consortium members

The data was ranked and put into Appendix K as Tables 32-35. Tables 32-35 show it is not easily possible to compare New Zealand with any of the countries directly across all five categories. This is mainly due to France and Germany having low remaining readable drive counts, and Australia having a large outlier of 78% for their company identifying drives which is inconsistent with previous years.

To make the comparisons easier to understand, the percentage is put first then the count for the country. In all comparisons where the percentages are equal, the country with the higher total drives analysed is listed first. The percentages are ranked as "best/most desirable" to "worst/least desirable" with an example in the case of unreadable hard drives, the lower the number of unreadable hard drives the better. New Zealand has been bolded to make the ranking clearer. The data for the individual countries was collected from the consortium 2009 results.

The two main areas of comparison therefore are the first two categories of unreadable and wiped. For unreadable drives by percentage, New Zealand is in the middle of the rankings, tied with Germany. New Zealand leads the wiped drives rankings.

As stated as the values in other categories being low, it is more appropriate to compare the consortium as a whole with New Zealand.

5.8.3 Summary of the primary objective

As Table 21 (on page 114) shows, the New Zealand results are consistent with the consortium results for 2009 overall. This suggests that New Zealand could either follow the recommendations from the consortium regarding educating users and the like, or New Zealand could be pro-active and create and test new recommendations for improving the wiping rates of second hand hard drives.

5.9 Secondary Research Objectives

There were two secondary research objectives for this thesis. The first was considered before the experiments started which was "Are certain types of company more likely to experience issues with data remanence than others?", and the second was "How accurate is the *Tableau* 'source drive may be blank' feature?" based on the use of the Spike Effect tool developed to assist automating the research.

5.9.1 Company Types

"Are certain types of company more likely to experience issues with data remanence than others?"

This secondary research objective was considered to be important as it was possible that one or more company types could have greater issues with data remanence.

As part of the documentation and record keeping for this research, when a company could be identified it was classified as a specific industry type and the data was also classified as how it was found. There is one caveat on the results that does not appear to have been reported by other researchers. That is, it is possible that a hard drive may have been owned by more than one company, or that the hard drive may have more than one company's data on it (a hard drive from an ISP or a company hosting web pages for example). For consistency, the drive would have been listed multiple times with each industry type as a separate entry.

There were a total of 19 drives with company identifying information on them. Each drive represented a single company so the caveat above was not relevant to this research. Unlike some of the other researchers into data remanence, such as Medlin and Cazier (2011) , specific counts of each type of readable data found (PowerPoint files and Word Documents rather than files that had identifying data in them) was not kept as it often ran into the hundreds of thousands and did not seem to add value to this research.

Table 24 on the next page uses *The Australian and New Zealand Standard Industrial Classification (ANZSIC) 2006* (Stats New Zealand, 2006) classification system (also used by Quinn (2010)) with the second column using the broad "Level 1" classifications. The full list of "Level 1" can be found in Appendix H. *ANZSIC* provides four levels of classification (Level 1 is a letter such as B for Mining, P for Education and Training, and then numbers are used to differential lower levels such as P80 for "Pre-school and School" or P81 for "Tertiary") but in certain cases identifying drives to the fourth or even third level could uniquely identify a company within New Zealand, or narrow it down to one of two companies. The third column values are based on the "File Category" classifications used within *FTK* to show examples of where the identifying information was found. The values (in alphabetical order) are "DA" databases, "DC" documents, "E" email, "F" folders or file system, "G" graphics, "O" other known types, "S" spreadsheets, "U" unknown type. "E" occurs most often as part of the methodology was to check emails first.

The fourth column was for notes that typically recorded where multiple drives had the same company data.

Table 24 Breakdown of Company Identifying Data found.

Drive name	First Level	File Category	Notes
IDE_53	Retail Trade	E	
IDE_58	Information Media and Telecommunications	S	Same as 56,61,66
IDE_59	Retail Trade	E	
IDE_60	Retail Trade	E,F	
IDE_61	Information Media and Telecommunications	E	Same as 56,58,66
IDE_66	Information Media and Telecommunications	F,S	Same as 56,58,61
IDE_69	Retail Trade	D,S	
IDE_70	Education and Training	D,E	
IDE_71	Retail Trade	E,S	
IDE_73	Manufacturing	DC,E	Same as 76,78
IDE_74	Education and Training	DC,S	
IDE_76	Manufacturing	DC,E,O	Same as 73,78
IDE_78	Manufacturing	DC,E,F	Same as 73,76
IDE_84	Administrative and Support Services	E,S	
IDE_87	Education and Training	E	
IDE_91	Education and Training	DC,E	
SCSI_23	Professional, Scientific and Technical Services	DC,E,S	
SCSI_24	Mining	DC,O	
SCSI_56	Information Media and Telecommunications	DC	Same as 58,61,66

One vendor (Company B) had four hard drives from the same large Telecommunications company. One vendor (One of the Trade Me vendors) had three hard drives from the same Manufacturing company. Table 25 presents the drives by the "first level" classification used by ANZSIC and used in Table 24, then presents the count of drives in that type, then the number of distinct companies (as derived from the Notes field in Table 24) and then the percentage of those companies.

Table 25 Table by Type of Industry

Classification	Number of drives	Number of distinct companies	% of distinct companies
Administrative and Support Services	1	1	7%
Education and Training	4	4	29%
Information Media and Telecommunications	4	1	7%
Manufacturing	3	1	7%
Mining	1	1	7%
Retail Trade	5	5	36%
Professional, Scientific and Technical Services	1	1	7%
Total	19	14	100%

Overall there were 14 unique companies represented. The four "Education and Training" hard drives could all be further classified as "School: Secondary Education" (P802200 using *ANZSIC*) whereas "Retail Trade" hard drives were diverse and included "Floor Coverings Retailing" and "Pharmaceutical, Cosmetic and Toiletry Goods Retailing" (G421200 and

G427100 using *ANZSIC* respectively) as examples. None of those four "Education and Training" drives were from the same secondary school and they were from three different vendors.

As Sutherland and Mee (2006) showed in their survey, British schools were often not properly disposing of their hardware, with a format and operating system re-installation being one of the main ways hardware was prepared for disposal. Therefore it is not a surprise that New Zealand showed a similar trend.

Otherwise, other researchers have not published industry based research results. The consortium publications do often mention specific instances of specific company types, but do not present overall statistics or observations into which company types are more at risk from data remanence. The most likely reason as shown from this research is the potential sample sizes needed to ensure an adequate number of company identifying hard drives (extrapolating New Zealand's results suggests over 500 hard drives examined to have 100 company identifying drives with a 10% margin error, assuming a "large" population base) and that certain company types may either not sell their hard drives (preferring to destroy the discs) or they follow government guidelines and properly sanitise their hard drives to the point that no information on them can be found. Therefore a different methodology and type of experiment may be required to answer this question, by both attempting to source drives specifically from companies (approaching schools to buy their hardware) or by using surveys or case studies of companies to determine how they do dispose of their hardware.

5.9.2 *Tableau* investigation

"How reliable is *Tableau*'s 'Source may be blank' function?"

The rationale for this research objective was the observation from using the Spike Effect tool that when *Tableau* stated "Source drive may be blank" the drive was not always blank. This in turn meant that the function may not be reliable and potentially meant spending more time than necessary on analysing some drives, and also underestimating the time required for the project for drives that did need to be analysed.

Of the 100 hard drives analysed, 39 were SCSI drives and therefore were not captured using *Tableau*. Of the remaining 61 drives (IDE and SATA) 13 were unreadable which left a total of 48 drives captured by *Tableau*.

A total of 22 drives out of 48 readable drives were reported as "source drive may be blank".

Table 26 breaks down these drives by drive type descriptions from the Spike Effect analysis (see Section 4.8.1 on pages 81-83) as well as one drive that could not be captured.

Table 26 Spike Effect output for "Source Drive may be Blank"

Type	Count	Percent	Notes
Capture incomplete	1	5%	*Tableau* initially started capturing but did not complete after multiple retries.
Practically Wiped	5	23%	These drives had less than 1 megabyte remaining.
Very Diverse Data (VDD)	6	27%	See discussion below.
Wiped	10	45%	These drives had no data remaining after 3 spikes were accounted for.
Total	22	100%	

Tableau stated one drive may have been blank based on its initial checking, however that drive could not be captured properly. It has been included in the results because of the initial checking, but as the notes state the drive itself was classified as "unreadable" as a proper forensic capture could not be completed. As the drive was unreadable, there was no Spike effect analysis.

15 of the drives were blank (or functionally blank) and *Tableau* was correct in labelling them as being blank.

Of the six drives that the Spike Effect classified as VDD, one drive (IDE_46) appeared to be reformatted, and data carving found no files. Using the flowchart (Figure 2 on page 65) the drive was therefore classified as wiped. IDE_46 did have a master boot record, but the rest of the drive was covered in what appeared to be random characters which suggest either a wiping program had been used, or the drive had been encrypted. It is also possible that the files were in a format that *FTK* could not recognize.

Of the remaining drives two were reformatted (IDE_74 and IDE_80) one of these contained identifying information (IDE_74) whereas none was found on the other (IDE_80). The

remaining three drives were clearly viewable which suggests they could be plugged into another computer and the data easily seen. Two of those three clearly viewable drives contained identifying data (IDE_59 and IDE_76) whereas the third drive (IDE_50) only contained an operating system on it. Data carving on IDE_50 did not find any identifying information.

Conversely, all drives that the Spike Effect reported as "wiped" were also correctly identified as "Source drive may be blank". For drives that were "practically wiped" *Tableau* did not consider IDE_39 or IDE_41 to be blank. It is likely that there were characters in the partition tables that did not trigger the "Source drive may be blank" function for those two drives.

Thus 27% of drives that *Tableau* identified as blank were not blank. 9% of the drives *Tableau* identified as blank actually contain identifying information on them.

Of the remaining 26 drives captured via *Tableau*, two drives were not classified as blank even though they were functionally blank as they each contained less than 3 kilobytes of data. It is true that they were not blank in the traditional sense, from the perspective of this research they are functionally blank. This represents 5% of the drives that would have required further investigation based on *Tableau*'s feedback, but not when the actual remaining data is considered.

Chapter 6: Conclusions

6.1 Introduction

This section addresses the primary and secondary research objectives as well as the research question raised during the experiments. The discussion will then consider the implications of those results for New Zealand, and then possible solutions.

Finally future work will be considered including experiments into "known state" drives as this thesis considered "unknown state" drives.

6.2 Primary Research Objective

The primary research objective was to "Determine a baseline of data remanence in New Zealand" and then use that to compare New Zealand data remanence with other countries.

As part of the methodology design for the experiments in this research the methodology for reporting results underwent a number of changes until the one finally used was agreed upon by the researcher and his supervisor. The changes considered how to deal with the total number of drives with identifying information (the consortium has not previously included this value and it makes understanding the overall level of data remanence difficult in some cases) and the relevance of reporting "illicit" material as the consortium research now covers a large geographic area and the jurisdictions of each country may have different definitions of what is illicit and what is not.

As for the results of the New Zealand data remanence into hard drives in the second hand market, overall they were directly comparable with the consortium 2009 (Jones, et al., 2010)when the consortium results were considered in aggregate. This is because the consortium members supplied results ranging from 17 drives for France, 39 and 42 for Germany and Australia respectively, 74 for North America and 174 for United Kingdom. There are two challenges with considering the consortium results individually, that is the sample size of 17 for France but also Australia's results which had a large outlier of 78% for company identifying information. Therefore, it made sense to consider the consortium as a whole, rather than individually.

The first observation is should New Zealand continue data remanence research if the results are similar enough to the consortium?

A practical way to reconsider that question is that because New Zealand appears to be in-sync with the consortium then New Zealand would make for a suitable and possibly interesting test bed for recommendations to people selling hard drives second hand that they should probably wipe their hard drives before selling them. This consideration makes further sense when considering that as an island nation reasonably far from its neighbours that New Zealand hard drives are less likely to be "contaminated" by other nation hard drives as seen in El Emam et al.'s research in Canada where hard drives formerly owned by United States citizens were found (El Emam, et al., 2007)

As the bulk of data remanence research into data sources other than hard drives has only been published in 2011, it is expected that other countries will incorporate those other devices (mobile phones and USB drives) into their research and New Zealand should also start investigating those devices. That would give a broader baseline for further measurements and analysis.

6.3 Secondary Research Objectives

6.3.1 Company Types

The first of the two secondary research objectives was to determine "Are certain types of company more likely to experience issues with data remanence than others?"

When this research question was initially considered, it was unclear what level of data remanence there would be on New Zealand hard drives.

Examining the overall data remanence from 2006-2009 for the consortium (Table 1 on page 44) suggests that if 100 New Zealand hard drives were examined there would be approximately 25 unreadable drives, 21 wiped drives and 54 readable drives with data remaining on them. Of those drives, approximately 21 drives (40%) would have company identifying information on them.

However when each table (Tables 28-31, Appendix I) was examined for the more extreme values the expected values become 10% unreadable, 33% wiped, 57% readable and 78% with company identifying information.

This suggests that it is possible there could have been between 9 and 47 hard drives with company identifying information in the New Zealand sample of 100 hard drives. After the experiments were conducted on New Zealand hard drives, a total of 19 hard drives had company identifying information on them. Of those drives, 14 were for different companies. 14 different companies are considered too low a sample size to make meaningful statements about however.

Therefore to fully answer this question, further investigation is required. As a baseline has now been created one approach would be to use El Emam et al.'s approach of purchasing and examining enough hard drives to produce statistically significant results regarding company identifying information. (El Emam, et al., 2007) The baseline would be useful as a planning tool since it would allow forecasting of the time and costs associated with the research.

As this research was non-targeted, a second approach would be to specifically target hard drives from specific company types such as high schools and determine the overall rates within those groups and then compare those groups. That research however has the potential issues raised by other researchers predominantly that vendors may be suspicious of requests for large numbers of drives of a specific type (El Emam, et al., 2007) and it may alert those vendors to better prepare their drives before disposing of them and it also may depend on when those company types were disposing of their hard drives.

The implications of this research objective do suggest that schools in particular need to take better care of their data when disposing of hard drives. This is in part due to schools having much more potential data on minors rather than adults and reasons why the school children may be absent from school (health or financial reasons) than small businesses.

Determining schools awareness levels of the GCSB manual (New Zealand Government Communications Security Bureau, 2010) will determine if further end user education for schools is required and then determine the best way of conducting that education.

For other companies, determining their level of understanding and commitment to the Privacy Act 1993 as determining if they have access to the GCSB manual will enable an effective strategy for improving end user education.

6.3.2 "How reliable is the *Tableau*'s 'Source may be blank' function?"

The other secondary research objective was to develop an analysis technique that could accurately and quickly determine if a hard drive was properly wiped or not. The "Spike

Effect" technique (and tool) attempted to achieve this. The tool is a byte histogram generator that can handle large individual evidence files (typically four gigabyte, but as large as 18 gigabyte within this research) as well as evidence files in an entire case (that ranged up to 80 gigabytes in this research). The tool then output the data into a text file which means it can be input easily into other programs for further analysis. As the initial results showed (Table 7 Comparison of Spike Effect program and *FTK* on page 86) that while the Spike Effect tool was reasonably quick, it was actually slower than *FTK* when processing cases that were wiped. Typically the Spike Effect program was run overnight which meant it could process multiple cases when processing time was not as important. In order to lower the processing time to less than the equivalent *FTK* time a variant of the Spike Effect tool was developed that processed random samples of the evidence files. The initial technique however was used to answer a related research objective of "How reliable is the *Tableau*'s 'Source may be blank' function?"

Once the experiments using the SATA and IDE drives started it became apparent that the *Tableau TD-1* device used for the forensic capturing of the hard drives would state "Source drive may be blank" when the drive itself was not blank, or it would not give the message when the drive was indeed blank. The makers of the *Tableau* do state in their manual that investigators should treat that statement with caution however.

This led to the research question of "How accurate is the *Tableau* 'Source drive may be blank' feature?"

Of 48 drives captured with *Tableau*, 22 drives were recorded as stating "Source drive may be blank". Of those 22 drives, 15 were blank and 6 drives had data on them and 1 could not be successfully captured. Of the other 26 drives that *Tableau* did not consider to be blank, 2 of them were functionally blank as they each had less than 3 kilobytes on data.

The implications are that drives that are blank when *Tableau* does not consider them to be blank will be captured and further analysed. Depending on the interface type of the drive, the amount of data on the drive (if the drive has been wiped with 0s those 0s will still be copied over) and therefore the size of the forensic image files and the analysis program, it may take hours or longer to determine that the drive was indeed blank. Multiplied over many drives this could end up taking up considerable time that could be better used in other ways.

Conversely and more importantly is if *Tableau* considers a drive to be blank when it does have data on it, the researcher may ignore the drive which in turn will alter the outcome of the results. Over a large sample that may alter the results by a few percent or smaller, but on a smaller sample it could produce very misleading results.

It must be noted that the following recommendation is based on one small function of the *Tableau* and that researchers and investigators should consider the other functions and features of a forensic copier before determining which one is appropriate for their purposes. This research did not consider other hardware solutions or determine if they had a similar function or not.

The initial recommendation would be to follow the *Tableau* manual's guideline of treating all hard drives as potentially having data on them (Tableau LLC, 2009) and continue the forensic investigation. The second recommendation would be to use a hardware or software solution (potentially live forensics if appropriate) to review the hard drive if possible either before forensic copying or after copying but before full analysis begins.

6.4 Future work

As this research determined the baseline for data remanence in New Zealand using "unknown state" drives there are a number of areas of future work.

The first main research area would be to implement the recommendations from this research and other researchers in the field of data remanence and determine if the data remanence rate changes over time due to those recommendations or from other reasons. A longitudinal study (three to five years) to compare year on year results within New Zealand and then with the consortium would be one way to achieve this. Experiments using targeted methodologies or by considering known state drives would also be possible research projects for future work.

The second main research area would be the issue of solid-state drives. It may be that all home users transition to solid state drives and that may prevent further data remanence research for second hand hard drives. There would be a number of research projects including estimating the date for when only Solid-State drives are sold in new and second hand computers. There could also be research into home user expectations of and attitudes towards solid-state drives, and if the concerns raised with regards to solid-state drive forensics is indeed relevant. This research could easily tie in with the first research area of a longitudinal study.

The third area of research would be to consider the other types of data storage that have data remanence issues not considered in this research. At the time the experiments were designed and conducted there had been very little published in the area of smart phones, USB flash drives and other devices for data remanence especially in the secondhand market place. The research that had been published did not include any research from New Zealand.

The fourth main research area is outlined in Section 4.8.5 and suggests that the Spike Effect tool has its own field of future work including improvements to processing time of the tool, and improving the heuristics to accurately determine if a hard drive is blank or not tools.

References

Adelstein. (2006). Diagnosing your system without killing it first. *Communications of the ACM, 49*(2), 4.

Anderson. (2003). You Don't Know Jack About Disks. *Queue*(June), 20-30.

Anderson, Durbin, & Salinger. (2008). Identity Theft. *Journal of Economic Perspectives, 22*(2), 23.

Anderson, Dykes, & Riedel. (2003). More than an interface — SCSI vs. ATA. *Proceedings of the 2nd Annual Conference on File and Storage Technology (FAST), March 2003*, 245-257.

Axelson. (2001). *USB Complete* (2nd ed.): Lakeview Research.

Chaerani, Clarke, & Bolan. (2011). *Information leakage through second hand USB flash drives within the United Kingdom.* Paper presented at the Proceedings of The 7th Australian Digital Forensics Conference, Perth, Western Australia.

Chen, Koutafy, & Zhang. (2009). Understanding Intrinsic Characteristics and System Implications of Flash Memory based Solid State Drives. *SIGMETRICS/Performance'09*.

Conroy. (2012). Sample Size A rough guide. Retrieved 21/9/2012, from http://www.beaumontethics.ie/docs/application/samplesizecalculation.pdf

Dicarlo. (2001). Summary of the Rules of Evidence. Retrieved 23 Feb 2012, from http://library.findlaw.com/2001/Jan/1/241488.html

Dufrasne, Letts, & Smith. (2004). Introducing IBM TotalStorage FAST EXP100 with SATA Disks. *IBM Redbook* Retrieved 24/11/2011, from http://www.redbooks.ibm.com/abstracts/redp3794.html?Open

El Emam, Neri, & Jonker. (2007). An evaluation of personal health information remnants in second-hand personal computer disk drives. *Journal of Medical Internet Research, 9*(3), 14.

Fragkos, Mee, Xynos, & Angelopoulou. (2006). *An empirical methodology derived from the analysis of information remaining on second hand hard disks.* Paper presented at the EC2ND 2006 : proceedings of the Second European Conference on Computer Network Defence, in conjunction with the First Workshop on Digital Forensics and Incident Analysis, University of Glamorgan Wales UK.

Garfinkel. (2007). Carving contiguous and fragmented files with fast object validation. *Digital Investigation, 4*, S2-S12.

Garfinkel. (2009). *Automating Disk Forensic Processing with SleuthKit, XML and Python.* Paper presented at the Systematic Approaches to Digital Forensic Engineering SADFE '09.

Garfinkel, & Shelat. (2003). Remembrance of data passed: A study of disk sanitization practices. *IEEE SECURITY & PRIVACY, 1*(1), 17-27.

Giannelli. (1983). Chain of custody and the handling of real evidence. *Am. Crim. L. Rev., 20*.

Glisson, Storer, Mayall, Moug, & Grispos. (2011). Electronic retention: what does your mobile phone reveal about you? *International journal of Information Security*, 337-349.

GNU C. ISO C Random Number Functions. Retrieved 6/11/2012, from http://www.gnu.org/software/libc/manual/html_node/ISO-Random.html

Gray. (2010). *Good Practice Guide for Identity Fraud Control*. Retrieved from http://www.dia.govt.nz/Pubforms.nsf/URL/Good_Practice_Guide.pdf/$file/Good_Practice_Guide.pdf.

Grispos, Storer, & Glisson. (2011). A comparison of forensic evidence recovery techniques for a windows mobile smart phone. *Digital Investigation, 8*, 23-36.

Gupta, Hoeschele, & Rogers. (2006). Hidden Disk Areas: HPA and DCO *International Journal of Digital Evidence, 5*(1), 1-8.

Gutmann. (1996). *Secure Deletion of Data from Magnetic and Solid-State Memory.* Paper presented at the Sixth USENIX Security Symposium.

Hooper, & Evans. (2010). The Value Congruence of Social Networking Services - a New Zealand Assessment of Ethical Information Handling. *The Electronic Journal Information Systems Evaluation, 13*(2), 12.

IBM. (2011, 1/15/2008). SCSI connector photos, specifications, and options. Retrieved 23/11/2011, from http://www-947.ibm.com/support/entry/portal/docdisplay?brand=5000008&lndocid=MIGR-4AQSCA

IEEE. (1990). 1149.1-1990 - IEEE Standard Test Access Port and Boundary-Scan Architecture. Retrieved 30/9/2012, from http://standards.ieee.org/findstds/standard/1149.1-1990.html

Johnson. (2009). *Am I Who I Say I Am? A Systems Analysis into Identity Fraud in New Zealand.* Auckland University of Technology AU.T, Auckland

Johnson, Calvert, & Raggett. (2009). *ICT in Schools survey 2009*: 2020 Communications Trust (NZ).

Jones. (2005). How much information do organizations throw away? *Computer Fraud & Security*, 6.

Jones. (2008). Industrial espionage in a hi-tech world. *Computer Fraud & Security*, 6.

Jones. (2009). Lessons not learned on data disposal. *Digital Investigation, 6*, 3-7.

Jones, Dardick, Davies, Sutherland, & Valli. (2009). The 2008 Analysis of Information remaining on disks offered for sale on the second hand market. *Journal of International Commercial Law and Technology, 4*(3), 162-175.

Jones, Valli, & Dabibi. (2009). *The 2009 Analysis of Information Remaining on USB Storage Devices Offered for Sale on the Second Hand Market.* Paper presented at the Proceedings of The 7th Australian Digital Forensics Conference, Perth, Western Australia.

Jones, Valli, Dardick, & Sutherland. (2006). The 2006 analysis of information remaining on disks offered for sale on the second hand market. *Journal of Digital Forensics, Security and Law, 1*(3), 23-36.

Jones, Valli, Dardick, & Sutherland. (2009). The 2007 analysis of information remaining on disks offered for sale on the second hand market. *Int. J. Liability and Scientific Enquiry, 2*(1), 16.

Jones, Valli, Dardick, Sutherland, Dabibi, & Davies. (2010). *The 2009 analysis of information remaining on disks offered for sale on the second hand market.* Paper presented at the 8th Australian Digital Forensics Conference, Edith Cowan University, Perth Western Australia,.

Joukov, Papaxenopoulos, & Zadok. (2006). *Secure Deletion Myths, Issues, and Solutions.* Paper presented at the Proceedings of the second ACM workshop on Storage security and survivability

Kahneman, & Tversky. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica, 47*(2), 263-292.

Krutov. (2011). IBM System x Server Disk Drive Interface Technology. *IBM Redbook* Retrieved 24/11/2011, from http://www.redbooks.ibm.com/abstracts/redp4791.html?Open

Kumar, Sofat, & Aggarwal. (2011). Identification and Analysis of hard disk drive in digital forensic. *International Journal of Computer Applications in Technology, 2*(5), 1674-1678.

Kwon, Lee, & Moon. (2006). Personal Computer Privacy: Analysis for Korean PC Users. In H. Yoshiura, K. Sakurai, K. Rannenberg, Y. Murayama & S. Kawamura (Eds.), *Advances in Information and Computer Security* (Vol. 4266, pp. 76-87): Springer Berlin / Heidelberg.

Langston. (2011). *Identity Theft Reported by Households, 2005-2010*. Retrieved from http://bjs.ojp.usdoj.gov/content/pub/pdf/itrh0510.pdf.

Leyden. (2009). Missile data, medical records found on discarded hard disks. Retrieved 30/11/2011, from http://www.theregister.co.uk/2009/05/07/data_destruction_survey/

Lyle. (2012). *Computer Forensics Tool testing Handbook*. Retrieved from http://www.cftt.nist.gov/CFTT-Booklet-Revised-02012012.pdf.

McKemmish. (1999). What is Forensic Computing? *Trends and Issues in Crime and Criminal Justice, 118*, 1-6.

Mead. (2006). Unique file identification in the National Software Reference Library. *Digital Investigation, 3*, 13.

Medlin, & Cazier. (2011). A Study of Hard Drive Forensics on Consumers' PCs: Data Recovery and Exploitation. *Journal of Management Policy and Practice, 12*(1), 27-36.

Min, & Nam. (2006). Current Trends in Flash Memory Technology. *ASP-DAC '06 Proceedings of the 2006 Asia and South Pacific Design Automation Conference*.

Mohamad, & Mat Deris. (2011). *Visualization Of Hard Disk Geometry And Master Boot Record.* Paper presented at the Universiti Malaysia Terengganu 10th International Annual Symposium, Universiti Malaysia Terengganu.

MSDN Microsoft. (2011). Windows and GPT FAQ. Retrieved 21/9/2012, from http://msdn.microsoft.com/en-us/windows/hardware/gg463525.aspx

National Computer Security Center. (1991). A Guide to Understanding Data Remanence in Automated Information Systems. Retrieved 2/10/2012, from http://www.fas.org/irp/nsa/rainbow/tg025-2.htm

New Zealand Department of Internal Affairs. (2011a). Evidence of Identity Standard. Retrieved 24/2/2012, from http://www.dia.govt.nz/diawebsite.nsf/wpg_URL/Resource-material-Evidence-of-Identity-Standard-Questions-and-Answers?OpenDocument

New Zealand Department of Internal Affairs. (2011b). Identity Theft. Retrieved 24/12/2012, from http://www.dia.govt.nz/Identity---What-is-identity-theft

New Zealand Films Videos and Publications Classification Act. (1993). New Zealand Films Videos and Publications Classification Act. Retrieved April 29 2011, from http://www.legislation.govt.nz/act/public/1993/0094/latest/DLM312895.html

New Zealand Government Communications Security Bureau. (2010). New Zealand Information Security Manual.

New Zealand Privacy Commission. (2011a). Principle 5. Retrieved 27/11/2011, from http://privacy.org.nz/storage-and-security-of-personal-information-principle-five/

New Zealand Privacy Commission. (2011b). Principle 11. Retrieved 19/9/2012, from http://privacy.org.nz/limits-on-disclosure-of-personal-information-principle-eleven/

NIST. (2006). Guidelines for Media Sanitization.

Owen, Thomas, & McPhee. (2010). *An Analysis of the Digital Forensic Examination of Mobile Phones.* Paper presented at the Fourth International Conference in Next Generation Mobile Applications, Services and Technologies.

Perl. (2003). Its not always about the money: Why the State Identity Theft Laws fail to adequately address criminal record identity theft. *The Journal of Criminal Law and Criminology, 94*(1), 41.

Piramanayagam. (2007). Perpendicular Recording media for hard disk drives. *Journal of Applied Physics, 102*(1), 1-23.

Quinn. (2010). 2010 New Zealand computer crime and security survey. Retrieved 24/2/2012, from http://otago.ourarchive.ac.nz/handle/10523/1455

Roberts, & Wolfe. (2011). *Data Remanence in New Zealand: 2011.* Paper presented at the Proceedings of The 9th Australian Digital Forensics Conference, Perth Western Australia.

Schmidt. (1997). *The SCSI Bus & IDE Interface* (Addison-Wesley, Trans. 2nd ed.): Addison-Wesley.

Schneier. (2008). The Psychology of Security. 27. Retrieved from http://www.schneier.com/essay-155.html

Sherman. (2006). *A digital forensic practitioner's guide to giving evidence in a court of law.* Paper presented at the Proceedings of the 4th Australian Digital Forensics Conference, Edith Cowan University Perth Western Australia.

Stats New Zealand. (2006). Industrial Classification (ANZSIC). Retrieved 21/9/2012, from http://www.stats.govt.nz/surveys_and_methods/methods/classifications-and-standards/classification-related-stats-standards/industrial-classification.aspx

Stuff.co.nz. (2011a). Dumped computers exploited by crims. Retrieved 30/11/2011, from http://www.stuff.co.nz/technology/gadgets/5725801/Dumped-computers-exploited-by-crims

Stuff.co.nz. (2011b). Old hard drives a fraud risk. Retrieved 30/11/2011, from http://www.stuff.co.nz/technology/gadgets/5793873/Old-hard-drives-a-fraud-risk

Surachit. (2007). Electro-mechanical hard drive diagram. Retrieved 25/2/2012, from http://en.wikipedia.org/wiki/File:Hard_drive-en.svg

Sutherland, Davies, Jones, & Blyth. (2010). *Zombie Hard Disks - Data from the Living Dead.* Paper presented at the 8th Australian Digital Forensics Conference, Perth Western Australia.

Sutherland, & Mee. (2006). *Data Disposal: how educated are your schools?* Paper presented at the The 5th European Conference on Information Warfare and Security, Helsinki, Finland.

Tableau LLC. (2009). TD1 Forensic Duplicator User Guide.

Thomas, & Tryfonas. (2007). Hard-drive Disposal and Identity Fraud. *IFIP International Federation for Information Processing New Approaches for Security, Privacy and Trust in Complex Environments, 232*, 461-466.

Thompson. (2005). MD5 collisions and the impact on computer forensics. *Digital investigation, 2*, 5.

Trade Me. (2011). I've paid and haven't received my goods or service. Retrieved April 27 2011, from http://www.trademe.co.nz/help/179/ive-paid-and-havent-received-my-goods-or-service

TradeIT. (2008). TradeIt auction sites. Retrieved april 27 2011, from http://www.tradeit.co.nz/nzsites.html

University of Otago. (2011). Ethical Practices in Research and Teaching Involving Human Participants. Retrieved 29 April 2011, from http://www.otago.ac.nz/administration/academiccommittees/otago015522.html

Valli. (2004). *Throwing out the Enterprise with the Hard Disk.* Paper presented at the 2nd Australian Computer, Information and Network Forensics Conference, Fremantle Western Australia.

Valli, & Jones. (2005). *A UK and Australian Study of Hard Disk Disposal.* Paper presented at the Proceedings of 3rd Australian Computer, Network & Information Forensics Conference, Perth, Western Australia.

Valli, & Woodward. (2007). *Oops they did it again: The 2007 Australian study of remnant data contained on 2nd hand hard disks.* Paper presented at the The 5th Australian Digital Forensics Conference, Edith Cowan University Mount Lawley Campus.

Valli, & Woodward. (2008). *The 2008 Australian study of remnant data contained on 2nd hand hard disks: the saga continues.* Paper presented at the 6th Australian Digital Forensics Conference, Perth Western Australia.

Weinstein. (1980). Unrealistic Optimism About Future Life Events. *Journal of Personality and Social Psychology, 39*(5), 806-820.

West. (2008). The Psychology of Security. *Communications of the ACM, 51*(4), 8.

Wikipedia. (2012). Wikipedia Stalking. Retrieved 25/2/2012, from http://en.wikipedia.org/wiki/Stalking

WikiQuotes. (2011). Cardinal Richelieu. Retrieved 19/2/2012, from http://en.wikiquote.org/wiki/Cardinal_Richelieu

Williamson, Apeldoorn, Cheam, & McDonald. (2006). *Forensic Analysis of the Contents of Nokia Mobile Phones.* Paper presented at the 4th Australian Digital Forensics Conference, Edith Cowan University, Perth Western Australia.

Yellow. (2011). Yellow Pages search for "Auction". Retrieved May 3rd 2011, from http://yellow.co.nz/search/dunedin/auction-1.html

Glossary

ATA: AT Attachment (ATA/ATAPI), the old name of Parallel ATA, an older interface for computer storage devices. It allows two devices on a cable.

Consortium: A group of data remanence researchers from Australia, France, Germany, United Kingdom and United States. The group came about based on research Valli (2004) from Australia, and Jones (2005) of the United Kingdom. The consortium is currently known as "Cyber Security Research Network".

Clearly viewable: This means the data on the hard drive can be easily examined or seen without the use of specialised tools. A typical user would be able to connect the hard drive to their machine and explore the hard drive easily.

Data Carving: Also known as file carving, it is the process of attempting to recover files or file fragments that have been deleted but not fully overwritten by the Operating system.

Data Remanence: Data that persists after efforts to remove it have been applied to a computer data storage device.

Degaussing: Degaussing as defined by NIST is exposing the magnetic media to a strong magnetic field in order to disrupt the recorded magnetic domains.

Deleting: Depending on the operating system, usually a file that has been selected for deletion is marked in such a way that the operating system knows it can reuse the space that the file has been allocated.

***Forensic Tool Kit* (*FTK*):** A toolkit for acquiring and analysing forensic computer evidence case files. Aspects of *FTK* have been tested by the Computer Forensic Tool Testing project which is associated with NIST.

Formatted: The consortium use the term formatted to mean "removing data by deletion, formatting or reinstalling an operating system" (Jones, Dardick, et al., 2009, pg. 165)

Hashing: Techniques that uses mathematical functions to produce a numeric value of files or a data source to determine if changes have been made, or if a file can be recognised.

IDE: Integrated Device Electronics, an earlier name for ATA

Known state: A device where the vendor has included a statement about their belief of what the state of data is on the device. Examples include "This hard drive has been wiped to DoD standards" or "This hard drive will be formatted before sale".

NIST: The American National Institute of Standards and Technology.

Purged: Specific category of data device sanitisation provided by NIST. It refers to removing data to the level that would likely defeat a laboratory attack (one that uses specialised equipment and training outside the normal operation of the data storage device). Degaussing is an example of purging.

Reformatted: This is the term used in this thesis to designate a hard drive that appears to have been formatted before sale. It means that a casual user would not see any data without using specialised tools. The consortium use the term "formatted" to mean this.

SATA: Serial ATA, an advanced form of ATA which provides a "Secure Delete" function amongst other improvements.

SCSI: Small Computer System Interface which provides a different style of interface to ATA. Advantages of SCSI it can allow up to 16 devices onto the cable.

Spike Effect: Both the name of a software program and an analysis technique designed for this thesis to help automate part of the forensic analysis process.

Tableau: *Tableau TD-1* is a hardware solution that allows for forensic duplication of hard drives and includes hashing functionality.

Unknown State: This is a device where the vendor has not explicitly made a statement regarding the state of data on the device. "As is where is" and "ex-lease computer" were deemed to be "unknown state" as it does not make explicit statements if the drive is readable or unreadable, or if the device has had some form of data wiping used on it or not.

USB: Universal Serial Bus which allows up to 127 devices to be connected.

Wiped: Under NIST's definitions wiped is akin to clearing which is defined as

> "Clearing information is a level of media sanitization that would protect the confidentiality of information against a robust keyboard attack. Simple deletion of

items would not suffice for clearing. Clearing must not allow information to be retrieved by data, disk, or file recovery utilities."

(NIST, 2006, pp. 15-16)

Appendices

Appendix A Price Comparison of Hard Drives.

The following is a brief comparison of hard drives prices in New Zealand on April 27, 2011

Data was retrieved from "Ascent" (www.ascent.co.nz) and "Pricespy" (www.pricespy.co.nz) with "Ascent" prices being used to check if "Pricespy" prices were up to date and accurate.

The Trade Me price is for the first instance of the drive found using the search features and the "Featured first" filter. The hard drives were selected based on the "Most Popular" listing on the "Ascent" website for the particular drive type and is illustrative of the approximate costs for the type of hard drive being described.

The cents have been removed from the price and the price has not been rounded to compensate for that.

- Hitachi Cinemastar 7K1000.C HCS721010CLA332 32MB 1TB

Ascent has this priced at $163.

Pricespy has 15 vendors listed with the price ranging from $142 to $188, with an outlier of $228 being the highest price listed. This is a SATA non Solid State Drive and it uses a SATA 3 Gb/s interface. This is a 3.5 inch drive.

- OCZ Vertex 2 Series SATA II 3.5" SSD 120GB

Ascent has this priced at $446.

Pricespy has 16 vendors with the price ranging from $408 to $480, with 1 company pricing it at $514. Trade Me did have 2 auctions for this particular drive, one from a vendor listed on "Pricespy" for $423 and one from a different vendor who did not appear to be listed on "Pricespy". As the $423 auction was "featured first" it was the price used in the following table. This is a SATA 3 Gb/s drive, and is 3.5 inches.

- OCZ Vertex 2 E Series SATA II 2.5" SSD 180GB

Ascent has this priced as $620

Pricespy lists 11 stores listed, ranging from $571 to $706. Again Ascent is roughly in the mid price range.

The OCZ Vertex 2 E Series SATA II 2.5" SSD 180GB was also listed on Trade Me as auction "auction-371042392". However it was an online retailer auction, as the listing both clearly stated it was "PB Technologies" and the price listed matched the price in the

"Pricespy" listings for "PB Technologies" of $588. This is a solid state drive, and uses a SATA 3 Gb/s interface. As the name states it is a 2.5 inch drive.

- Seagate Momentus XT ST95005620AS 32MB 500GB

Ascent has this listed as $188 but out of stock. Pricespy has 44 vendors listed, with the price ranging from $167 to $240, with an outlier of $280. The Trade Me price in auction "auction-370504850" listed the price as a "Buy now" of $225. There was no indication that it was one of the vendors listed by Pricespy. It uses a SATA 3 Gb/s interface and was listed on Trade Me as being a "laptop" hard drive as it is a 2.5 inch drive.

Table 27 Comparison of Hard Drive prices

Name	Type	Form Factor	Interface	Size in Gb	Ascent Price	Pricespy Range	Trade Me Price
Hitachi Cinemastar	Non Solid State	3.5	SATA 3 Gb/s	1000	$163	$142 - $188	*
OCZ Vertex	Solid State	3.5	SATA 3 Gb/s	120	$446	$408 - $480	$423
OCZ Vertex	Solid State	2.5	SATA 3 Gb/s	180	$620	$571 - $706	$588
Seagate Momentus	Hybrid	2.5	SATA 3 Gb/s	4 : 500 **	$188 ***	$167 - $240	$225

* No auction was found that matched this particular hard drive.

** As this is a hybrid drive, it has 4 Gb of Solid State Drive, and 500 Gb of Non Solid State drive.

*** Out of stock at the time of writing.

Appendix B Output of Spike Effect on Drive SCSI_03

Array	0	=	252	and	as	a	percent	0.00%
Array	1	=	6	and	as	a	percent	0.00%
Array	2	=	3	and	as	a	percent	0.00%
Array	4	=	2	and	as	a	percent	0.00%
Array	5	=	1	and	as	a	percent	0.00%
Array	8	=	1	and	as	a	percent	0.00%
Array	10	=	1	and	as	a	percent	0.00%
Array	12	=	1	and	as	a	percent	0.00%
Array	14	=	1	and	as	a	percent	0.00%
Array	16	=	5	and	as	a	percent	0.00%
Array	19	=	3	and	as	a	percent	0.00%
Array	24	=	1	and	as	a	percent	0.00%
Array	30	=	1	and	as	a	percent	0.00%
Array	31	=	2	and	as	a	percent	0.00%
Array	32	=	10	and	as	a	percent	0.00%
Array	33	=	2	and	as	a	percent	0.00%
Array	34	=	1	and	as	a	percent	0.00%
Array	38	=	1	and	as	a	percent	0.00%
Array	39	=	1	and	as	a	percent	0.00%
Array	44	=	1	and	as	a	percent	0.00%
Array	48	=	827127838	and	as	a	percent	**4.54%**
Array	49	=	1654535587	and	as	a	percent	**9.09%**
Array	50	=	827127839	and	as	a	percent	**4.54%**
Array	51	=	827127838	and	as	a	percent	**4.54%**
Array	52	=	827407746	and	as	a	percent	**4.55%**
Array	53	=	827407747	and	as	a	percent	**4.55%**
Array	54	=	827407745	and	as	a	percent	**4.55%**
Array	55	=	827407745	and	as	a	percent	**4.55%**
Array	56	=	827407744	and	as	a	percent	**4.55%**
Array	57	=	827407745	and	as	a	percent	**4.55%**
Array	59	=	1	and	as	a	percent	0.00%
Array	60	=	1	and	as	a	percent	0.00%
Array	61	=	1	and	as	a	percent	0.00%
Array	62	=	1	and	as	a	percent	0.00%
Array	63	=	2	and	as	a	percent	0.00%
Array	65	=	827407748	and	as	a	percent	**4.55%**
Array	66	=	1654535584	and	as	a	percent	**9.09%**
Array	67	=	1654535583	and	as	a	percent	**9.09%**
Array	68	=	1654535582	and	as	a	percent	**9.09%**
Array	69	=	1654535585	and	as	a	percent	**9.09%**
Array	70	=	1654815492	and	as	a	percent	**9.09%**
Array	77	=	1	and	as	a	percent	0.00%
Array	85	=	4	and	as	a	percent	0.00%
Array	86	=	1	and	as	a	percent	0.00%

Array	94	=	1	and	as	a	percent	0.00%
Array	97	=	6	and	as	a	percent	0.00%
Array	98	=	1	and	as	a	percent	0.00%
Array	99	=	1	and	as	a	percent	0.00%
Array	100	=	4	and	as	a	percent	0.00%
Array	101	=	8	and	as	a	percent	0.00%
Array	102	=	2	and	as	a	percent	0.00%
Array	103	=	2	and	as	a	percent	0.00%
Array	104	=	1	and	as	a	percent	0.00%
Array	105	=	9	and	as	a	percent	0.00%
Array	108	=	1	and	as	a	percent	0.00%
Array	110	=	6	and	as	a	percent	0.00%
Array	111	=	5	and	as	a	percent	0.00%
Array	112	=	2	and	as	a	percent	0.00%
Array	113	=	1	and	as	a	percent	0.00%
Array	114	=	12	and	as	a	percent	0.00%
Array	115	=	1	and	as	a	percent	0.00%
Array	116	=	7	and	as	a	percent	0.00%
Array	117	=	7	and	as	a	percent	0.00%
Array	118	=	2	and	as	a	percent	0.00%
Array	119	=	1	and	as	a	percent	0.00%
Array	124	=	8	and	as	a	percent	0.00%
Array	125	=	2	and	as	a	percent	0.00%
Array	128	=	2	and	as	a	percent	0.00%
Array	129	=	4	and	as	a	percent	0.00%
Array	130	=	1	and	as	a	percent	0.00%
Array	137	=	2	and	as	a	percent	0.00%
Array	138	=	1	and	as	a	percent	0.00%
Array	139	=	3	and	as	a	percent	0.00%
Array	141	=	2	and	as	a	percent	0.00%
Array	142	=	6	and	as	a	percent	0.00%
Array	163	=	3	and	as	a	percent	0.00%
Array	165	=	1	and	as	a	percent	0.00%
Array	167	=	1	and	as	a	percent	0.00%
Array	170	=	4	and	as	a	percent	0.00%
Array	172	=	1	and	as	a	percent	0.00%
Array	180	=	3	and	as	a	percent	0.00%
Array	182	=	1	and	as	a	percent	0.00%
Array	184	=	3	and	as	a	percent	0.00%
Array	185	=	1	and	as	a	percent	0.00%
Array	187	=	2	and	as	a	percent	0.00%
Array	189	=	1	and	as	a	percent	0.00%
Array	190	=	3	and	as	a	percent	0.00%
Array	192	=	4	and	as	a	percent	0.00%
Array	193	=	1	and	as	a	percent	0.00%
Array	195	=	3	and	as	a	percent	0.00%

Array	199	=	2	and	as	a	percent	0.00%
Array	205	=	5	and	as	a	percent	0.00%
Array	208	=	3	and	as	a	percent	0.00%
Array	213	=	1	and	as	a	percent	0.00%
Array	215	=	1	and	as	a	percent	0.00%
Array	216	=	3	and	as	a	percent	0.00%
Array	219	=	1	and	as	a	percent	0.00%
Array	224	=	3	and	as	a	percent	0.00%
Array	232	=	4	and	as	a	percent	0.00%
Array	233	=	3	and	as	a	percent	0.00%
Array	234	=	2	and	as	a	percent	0.00%
Array	235	=	1	and	as	a	percent	0.00%
Array	238	=	1	and	as	a	percent	0.00%
Array	239	=	1	and	as	a	percent	0.00%
Array	241	=	1	and	as	a	percent	0.00%
Array	243	=	2	and	as	a	percent	0.00%
Array	246	=	8194	and	as	a	percent	0.00%
Array	248	=	1	and	as	a	percent	0.00%
Array	250	=	1	and	as	a	percent	0.00%
Array	251	=	2	and	as	a	percent	0.00%
Array	252	=	1	and	as	a	percent	0.00%
Array	253	=	1	and	as	a	percent	0.00%
Array	254	=	3	and	as	a	percent	0.00%
Array	255	=	4	and	as	a	percent	0.00%

There were a total of 18200739840 values in the array

This was composed of 114 different characters read into the array.

* Certain percentages have been bolded to make them clearer to read.

Appendix C Full results of the experiments

The following present the full results of the experiments, ordered by source (Company A, Company B, Trade Me). The results are generally in numeric order, except where a group of drives share the same characteristics such as SCSI_10 appears before SCSI_05 because SCSI_10 shares the same characteristics as SCSI_02 which was discussed before SCSI_05. The terms "clearly viewable" and "reformatted" are explained in the glossary but briefly refer to if tools were needed to find and view the data or not. The pseudonyms "Jane Doe" and "John Doe" has been used when examples of individual identifying information were needed. Where a company name was needed "fake_company_name" was used.

Results from Company A : Drives 1 to 50

Brief results:

Company A supplied 50 hard drives ranging from 4 gigabytes to 40 gigabytes per hard drive.

One of the drives (SCSI_17) had individual identifying information only. SCSI_23 had company identifying information only. SCSI_24 had both company and individual identifying information on it. All three drives had been reformatted and data carving was required to find the information.

Of the drives that did not have identifying information on them, two drives had not been reformatted, and contained Operating System and other files. As per the methodology when no identifying information was found in those visible files, data carving was then used to find any deleted files. No identifying information was found in the files that data carving managed to recreate.

Further analysis:

Company A sent the drives in three groups. The first group consisted of 10 SCSI drives and the remaining two groups had 20 drives each.

The first hard drive from the first group to be analysed was classified as "unreadable" by the testing system. The remaining nine drives were all readable and attempts were then made to try to forensically copy SCSI_01 again. As those attempts failed on SCSI_01, SCSI_01 was classified as unreadable.

The 9 readable drives were 18G drives each, with the exception of SCSI_07 which was a 9G drive.

The "Spike Effect" program was then run on the forensic copies of the 9 remaining drives. It was observed that 6 drives shared identical characteristics (SCSI_03, SCSI_04, SCSI_06, SCSI_07, SCSI_08, and SCSI_09) from the output of the Spike Effect having been run on their forensic copies and they were examined first. Appendix B contains the full "Spike Effect" output from SCSI_03 where the array entries had a value greater than 0. Of note, only 114 of the 256 possible array entries had values greater than 0. The results showed the 3 largest spikes each being 9.09% each. Further investigation of the "Spike Effect" output showed that there were 6 spikes of approximately 9.09% each and 10 spikes of approximately 4.54% each. In total this accounted for at least 99.99% of the data on the drive.

The expected outcome based on the Spike Effect output was that *FTK* would be unable to carve any files for SCSI_03. When *FTK* attempted data carving for SCSI_03 no files were recovered. For consistency drives SCSI_04, SCSI_06, SCSI_07, SCSI_08, and SCSI_09 were analysed as well. The results for those drives were functionally the same as SCSI_03. The pattern used was identical, and the Master Boot Records were similar. These drives were then classified as "Special Case" and as "Wiped" as per the flowchart.

SCSI_02 was an 18 GB hard drive and had a very different "Spike Effect" output. The most obvious difference was that all 256 array entries had values greater than 0 which means that there is potentially data on the drive (either clearly viewable or found after data carving). After the three largest spikes were taken into consideration there was still 87% of the hard drive or 15 GB of data left. The first spike accounted for approximately 10% of the drive and was the array value 0 or "Null".

When the forensic copy was loaded into *FTK* the drive appeared reformatted and was noted as such. Data carving did return a large number of files, however they were Operating System related and none yielded any form of meaningful identifying information. This suggests that the former owner was using it as a server and had reformatted the drive before disposing of it.

SCSI_10 also had similar "Spike Effect" values to SCSI_02. When analysed it appeared to be reformatted and data carving was then used. Data carving did return a large number of files, however they Operating System related and none yielded any form of meaningful identifying information.

SCSI_05 also had all 256 array values being greater than 0. In contrast to SCSI_02 the three largest spikes accounted for only 2% of the total drive. This presented a consistent spread of values across the entire drive which means there could be information that could be successfully data carved, or alternatively that a pattern had been used to wipe the drive, possibly some form of encryption had been used to wipe the drive. Examining the drive with *FTK* however found the drive was reformatted. On further examination it appeared there was a random pattern had been selected and then repeated across the hard drive, and that pattern consisted of all 256 characters. The Master Boot Record was still viable and was very similar to Figure 7 (on page 83) without a /P pattern associated which suggests that Scrub3.exe will generate a random pattern if no specific pattern is selected.

The initial observation of reformatted was changed to and recorded as "Wiped" for SCSI_05 as per Figure 2 (on page 65) and it was classified as "Special Case".

The second shipment of drives contained 20 hard drives. These ranged in size from 4 GB to 18 GB and were a mix of SCSI and SCSI-SCA. SCSI-SCA drives were recorded as SCSI for naming convention purposes. Therefore the names started with SCSI_11 and went to SCSI_30. This again required some minor modifications to the testing system to allow reading of SCSI-SCA drives.

In the second shipment from Company A, the first drive that was attempted to be read (SCSI_11) was unreadable. Subsequently SCSI_12, SCSI_13, and SCSI_14 were also unreadable which raised concerns that the drives were damaged in transit. Drives SCSI_15 to SCSI_26 were readable however and no other explanation could be determined other than "luck of the draw" that four of five unreadable drives happened to be the first four drives picked out of the box. The 5th unreadable drive from this shipment was SCSI_27.

After using the Spike Effect program on the readable drives in that shipment it was noted that SCSI_16, SCSI_20, SCSI_21, SCSI_22, SCSI_30 all had 99.98% or more of data in the largest Spike alone, and all 5 had less than 300 kilobytes of data left on each drive after the three largest spikes were accounted for.

Examining SCSI_16 and SCSI_20 in *FTK* showed that as expected, one character had been used to wipe the bulk of the hard drive, and data carving was used despite it being likely nothing would be found. No files were able to be carved from SCSI_16 and SCSI_20. This

meant the drives were classified as wiped. SCSI_21, SCSI_22 and SCSI_30 were recorded as being as well.

The Spike Effect for the remaining hard drives in this shipment all had values greater than 0 for all 256 array entries and therefore required further investigation with *FTK*.

SCSI_17 had approximately 5 GB of data left, and appeared to be reformatted. After data carving a large number of documents and emails were found. The emails were examined first and they suggested it was an individual user as the emails discussed the sender "submitting their work and going on holiday". Although the emails dated from 2004 this drive was used an example of identifying information as it contained details of the person's name, address, phone number and then other personal information not commonly found elsewhere.

SCSI_19, SCSI_26, SCSI_28 were all reformatted and data carving was used. As a group they contained a number of agriculture themed images and documents such as maps and overlays of beehives, sick animals and other information that may have been research data or educational material. All three drives appeared to be similar in make, model, capacity and from the data found on them, it is possible they may have been from the same raid array, or from the same original source. Despite in-depth examination of the files it was unclear if the drives belonged to a farmer, a (university) student, a lecturer/researcher or a government agency and as such could not be classified as having identifying information.

SCSI_23 was reformatted, and data carving revealed a large number of documents, some emails and a few images. The images were of importance as one was a fax which included a company name and other relevant details, and other images included scans of legal documents. A tax invoice was also found with the company name, employee name, person being invoiced, the amount being invoiced, and the date which was in the year 2000. Comparing the information found it was reasonably clear which company the hard drive came from, and there was a reasonable amount of company identifying data on the drive.

SCSI_24 was also reformatted and data carving found a large amount of data. Amongst documents found was an offer of employment from a company to an individual. Another document had the company name and discussed a purchasing agreement. Amongst the html files was a "telephone directory / contact list" for the same named company. There were enough details from other documents to reasonably assume the person who signed the letter

making the offer of employment was also the user of the computer. The drive therefore was classified as both company identifying and individual identifying.

SCSI_15, SCSI_18, SCSI_25, SCSI_27, and SCSI_29 were all reformatted and data carving produced files that did not have any identifying information in them.

The third group of hard drives contained a mixture of IDE, SATA and SCSI drives. The sizes ranged from 6 GB to 20 GB for the non-SATA drives with the three SATA drives being 40 GB each. SATA drives potentially transfer data faster than IDE drives so while the drives were larger, the improved speed meant they were captured in approximately the same amount of time as the smaller drives.

Three drives in this group were unreadable, those being SCSI_34, SATA_37, IDE_43.

The Spike Effect output for IDE_44, IDE_45, and IDE_49 all had one spike of 100% for each of them. This meant they were classified as "wiped".

10 drives (SCSI_31, SCSI_32, SCSI_33, SATA_36, IDE_39, IDE_40, IDE_41, IDE_42, IDE_47, and IDE_48) had less than one megabyte remaining on each drive and were classified as "wiped" based on the Spike Effect "practically wiped" category.

Therefore there were three remaining drives that potentially had data on them SCSI_35, SATA_38 and IDE_50.

SCSI_35 was clearly viewable and appeared to be a server, based on the operating system found on it, and other files. None of the files had any identifying information so data carving was used. No files were returned from data carving. There are some possible explanations for this such as the former owners had wiped the drive and then put a new operating system on to the drive before disposing of it, or that it was a test server that was hardly used, or that the former owners used a particularly good data removal tool and targeted data they considered sensitive or requiring removal. There was not enough information to determine which of those explanations was correct or if there was another explanation.

SATA_38 had 2 GB remaining of 40 GB. It also was clearly viewable and appeared to have "Windows XP Pro" on it based on the boot.ini file on the drive. There were no spreadsheets or databases found, but some documents and images. These did not yield any identifying information and while data carving did produce more files, those too did not have any identifying information in them.

IDE_50 has only 859 megabytes remaining out of 20 GB or approximately 4%. As the largest spike covered approximately 19 GB (95%) of the drive this suggested that either the drive had been wiped and some data had been written back to it such as a small operating system or that the hard drive had been partitioned and only one partition had been wiped. A third alternative is that the data in the 19 GB was never used and the character found by the Spike Effect was the original character on the hard drive. *FTK* found no emails and a small number of documents. Data carving only returned approximately 50 html and document files which also did not have any identifying information in them.

Overall the drives from Company A covered a range of sizes from 4G to 40G, and covered the three main interface types examined in this research of SCSI, SATA and IDE. Only 3 drives had identifying information on them.

Results from Company B : Drives 51 to 69

All 19 drives were shipped in one group. 15 were IDE, and 4 were SCSI.

Nine of the drives were unreadable. They were IDE_51, IDE_52, IDE_57, IDE_62, IDE_63, IDE_64, IDE_67, IDE_68 and SCSI_55. Those drives differed in brand, make, size and apparent age. The reasons for these drives being unreadable included "bad sectors", one drive made a strange metal pinging sound and wouldn't capture, and *Tableau* not detecting the source drives for any reason in some cases. Typically when *Tableau* would not detect a source drive it meant the jumper settings were not correct, however as per the methodology, the jumpers were changed through all configurations and the drives would still not be detected by *Tableau*. Given the age of some of the other drives it is also likely that the age of these drives may have contributed to them being unreadable and damage in transit is possible.

The Spike Effect output for each readable hard drive indicated "very diverse data" on the drives. Of the 10 readable drives, IDE_53, IDE_59, IDE_60, IDE_61, IDE_66, IDE_69 and SCSI_65 were clearly viewable, and each had an operating system with data files on them. The remaining three drives IDE_58, SCSI_54 and SCSI_58 were reformatted.

IDE_53 had a list of users using the full first name, and initial of the person's last name i.e. John_D which was found by *FTK*'s "disk view" function. Other data included invoices and emails which matched the user names, and a company name.

SCSI_54, SCSI_55 and SCSI_56 appeared to be part of a raid array from the external markings on the drives themselves however as per the methodology the drives were treated as separate drives.

Both SCSI_54 and SCSI_56 had been reformatted, and after data carving SCSI_54 returned very few files. Those files were not identifying in any way. The files that were recovered on SCSI_56 however included a number of resumes/C.Vs and other personally identifying data relating to two people. The company identifying information included a list of computers with their network names, user names and passwords in clear text. Typically this would be of concern as it could easily allow for corporate espionage and also if the users had selected their own passwords it may make guessing the new passwords easier. This data was from 1995 and therefore unlikely to be of use today. SCSI_55 was unreadable.

IDE_58 was a small laptop hard drive. By using a converter it was possible to connect it to *Tableau* for imaging and therefore classified as being an IDE drive. The drive was reformatted, and after data carving it appeared to be a staff member's laptop for a large New Zealand company. This is based on a number of spreadsheets with the company's name and various elements that contextually related to the nature of the business of the company. The details also appear similar to those found on IDE_66. Nothing about the staff member was found however.

IDE_59 was a 40 GB hard drive, and was a drive that *Tableau* stated "Source drive may be blank". The "Spike Effect" for this drive showed all 256 characters represented and approximately 87% for the first spike, and the other 2 spikes being 0.5% and 0.25% respectively. This meant there was potentially 5 GB of data left on the drive. The hard drive was clearly viewable and appeared to be a company drive as the "disk view" showed a list of user names in the format of full first name and full last name i.e. John_Doe. When emails were examined there were a number of business quotes to customers. Examining one of the user directories however found one of the most interesting pieces of data in this research.

The user had signed up for a virtual sports betting site which appeared to be legal within New Zealand as no real money was used. The web form had asked for (and stored) the following values "User name", "First name" "Last name", "Street address". "Email address" "Phone number" "Cell phone number" and "Date of birth". There was a "password" field however this did not store the given password and no attempts were made to recover the password.

The user had put in an email address in the format of Johndoe@fake_company_address.co.nz which was consistent with the company named elsewhere on the hard drive. The user's name was also consistent with data found on the hard drive. The other details were all consistent with street addresses and telephone area codes in use in New Zealand. Not only was the data of concern given a credible date of birth was provided, other information found on the hard drive (email dates, web page information as examples) showed the data appeared to date from 2009.

IDE_60 was a straight forward to analyse hard drive. It was clearly viewable, and had over 300 emails. Those emails were predominantly from the same sender, that being a manager within the company that was identified. The emails typically were about marketing initiatives, budgets and sales figures as well as praising certain staff for meeting sales criteria and winning internal company prizes. Additional emails were specific orders placed by customers to the particular staff member who used the computer, which included the customer names and physical addresses. Other emails were of a personal nature with the staff member sending emails to their children's schools with sick notes and the like.

IDE_61 was clearly viewable and had a number of emails. Most emails were in a "complaints" email folder with people sending complaints about products to the company. As such occasionally other contact details for those people were found from the emails. Spreadsheets also contained a variety of information about suppliers and inventory lists. Other information was of an individual identifying nature to the main user of the machine.

IDE_66 was an older laptop hard drive similar to IDE_58. The hard drive made a number of unusual clicking noises when being imaged which suggested it would be unreadable, however *Tableau* was able capture the data. The hard drive was clearly viewable and contained data with references to the same company as IDE_58. The main difference with this drive from IDE_58 is there were also clear references to Novell Netware 4 and appears it was a remote client that connected to a Netware server.

IDE_69 was clearly viewable and had an operating system and other files on it. After examining the files there was nothing identifying so the data carving was used. The files found from data carving contained psychometric tests for three employees (names and then relative attributes in a variety of situations such if the person being tested works well in group situations, has leadership potential, is honest and such like). Of note with the psychometric test however was the word "Confidential" at the bottom of each page, and references to the

1993 New Zealand Privacy Act such that the employers needed to concern themselves with the proper storage and deletion of the tests.

Other files on the hard drive included spreadsheets, budgets and price quotes. The data therefore was company identifying, and due to the psychometric testing and other data found, individual identifying.

Results from Trade Me : Drives 70 to 100

A total of 31 drives were sourced from auction sites. Initially 29 had been purchased however when the drives were shipped (in four groups) two extra drives had been added by the vendors.

The first group was 10 IDE drives. One drive (IDE_77) was unreadable. After running the Spike Effect program two drives (IDE_72, 79) appeared to be wiped and the remaining drives had very diverse data on them.

IDE_70 was an interesting drive. The drive was clearly viewable and during the initial *FTK* analysis the onboard anti-virus program on the tower used for analysis flagged five documents as being possible malware infections and quarantined those files. Those files were not investigated further. The data that were visible on the drive did not have identifying information but after data carving it was clear from the documents and spreadsheets that the hard drive had belonged to a high school and therefore recorded as "company identifying".

IDE_71 would not initially copy when connected to *Tableau*. The jumper settings were changed multiple times and eventually it was readable. The drive was clearly viewable and had over 800 emails. These were clearly company identifying as they were in the format of jane_doe@fake_company_name.co.nz to various other staff in the company. Most of the emails were either sales data or upcoming promotions, or congratulations to specific staff about meeting specific business goals. Examining the documents and spreadsheets also had more in-depth promotional information such as "product X has 33% mark-up, for next week, lower that to 10% mark-up". It is possible that this data could be potentially useful to rival companies however the last email was sent in October 2008.

IDE_73, IDE_76, and IDE_78 were all clearly viewable and were from the same company. IDE_73 had information up to 2008 and IDE_78 had information up to January 2011. There were a number of emails which were from the company to customers both nationally and internationally. These emails had bids and tender information and checking other documents

showed the costings and potential profit for those bids. While other companies in the same industry potentially have similar formulas it is possible some of the information is either commercially sensitive or could give a rival company a competitive advantage. Each drive was recorded separately as having company identifying information.

IDE_78 had a number of issues when *FTK* was attempt to process the forensic copy case files which meant it could not be fully processed. The processing had to manually stopped once *FTK* had processed approximately 280 000 files. This was after approximately 20 hours of processing time. It should be noted *FTK* had successfully processed drives that had over 400 000 files as part of this research so the volume alone was not the issue. Multiple attempts at the processing had been done before the decision was made to stop the processing at the 280 000 file point. It appears to have been a server machine and whilst it definitely had company identifying, it was not possible to find individual identifying information in the 280 000 files processed.

IDE_74 was a reformatted hard drive. Data carving produced a large number of spreadsheets. Examining those spreadsheets revealed a number of sheets with a high school's name in it. This high-school was not the same as IDE_70. The data revealed income and expenses as well as boarding fees and other fees. There were also rosters for staff members. Other than the names of those staff members (equivalent as being on a school website or in the telephone directory), there was no other individual identifying information and therefore the drive was classified as company identifying only.

IDE_75 had a file that caused *FTK* to stop processing in a similar way to IDE_78. The drive had not been reformatted. Due to the large number of files that had been processed however it was possible to determine the drive had individually identifying information. Examining the disk view showed a "guest" account and two users with the same last name on it. Examining the files further showed "Jane Doe" was selling cosmetics from home. This was not considered to be company identifying as per the criteria used on page 63 (as this appeared to be a "sole trader") however and the spreadsheets did not contain any business related data. Emails suggested "Jane Doe" had recently had a child and pictures had been attached and sent out. The drive was therefore classified as "individually identified" but not "company identifying".

Overall, there not appear to be any connection between the owners of IDE_75 and drives IDE_73, 76, 78. Other than Drive 71 and 74 both being high schools there appeared to be no

connection between them either. Of note is that all seven readable drives had amounts of identifying information, and only one of those drives had been reformatted.

The second group of auction drives consisted of 6 IDE drives. This auction lot contained predominantly 20G to 40G drives and the Spike Effect for all 6 showed the drives to have very diverse data.

IDE_80 was a reformatted drive. After data carving, the drive appeared to have previously hosted a web server or been a test bed for a web server for a high school. This is based on the large number of html files, their organization and internal linking and the large number of high school themed graphics files associated with the pages. The files and the graphics however did not identify which high school it was however, and also did not identify individuals in any way. Some "Google search" web pages were also found and they had copyright dates of 2003 and 2004 which suggest the hard drive may have been at least seven years old.

IDE_81 was clearly viewable and had two partitions on it. One of those partitions had the Windows 98 operating system on it and had computer games installed that were at least six years old. The second partition appeared to be for storage as it did not have any operating system files on it. Disk view did not return any user names, and viewing the web browser cache found many graphics which were mainly Australian themed images, none of which were illicit. The computer games and lack of other information suggest it was a home user computer and was not identifying in any way.

IDE_82 was another clearly viewable drive. There were over 300 emails, mostly from one person discussing various events including a wedding. This was further confirmed by the inclusion of wedding images. The sender also appeared to in charge of a sports club and sent out regular emails to other members about upcoming events. This was not considered to be company identifying as per the criteria used on page 63. The decision was made not to classify this as a company identifying machine but it was classified as individually identifying.

IDE_83 was a clearly viewable drive. Examining the files did not find anything identifying so data carving was used. The results of the data carving did not find anything identifying.

Using the classification flowchart (Figure 2 on page 65), the drive was classified as "OS/F" to represent it had an operating system and files but was not wiped.

IDE_84 was another clearly viewable drive. Much like the IDE_78, this drive would stop processing consistently after 90 minutes when using *FTK*. The decision therefore was made to stop processing at the 89 minute mark and analyse what was found. As there was a lot of company identifying and individually identifying data found from what *FTK* was already able to process it was not considered as concerning that the whole drive could not be fully processed.

IDE_85 was clearly viewable. Data carving found a large number of generic web pages and images and no identifying information could be determined from those pages.

The third group of drives were all IDE drives. Two of the drives were unreadable (IDE_92 and IDE_93) and none were wiped.

IDE_86 was a reformatted hard drive. Data carving found a number of web pages that had interesting individual information. Similar to IDE_59 that had a web form with user information; this hard drive had an online Golf tracking system where players are encouraged to post their games. They were asked to put in player real life names, the course played, the time and date, and the scores for each player. This was in part to determine who the better players were and to track playing over time.

The other main web form that the owner of the hard drive interacted with was an Australian lottery syndicate which initially appeared to let users purchase tickets for a number of Australian lotteries. On subsequent pages the user was encouraged to sign up friends and family, and the initial user would then get discounts and/or benefits from those signing up. This initially appeared to either be multi-level marketing which is legal in Australia and New Zealand (under certain conditions) or a pyramid scheme (which is illegal in both countries). Because it was unclear if the syndicate was legal or not, this researcher decided to investigate it further. This in turn led to a “scam busters” style website and the syndicate was discussed at length in their sub forums. The consensus was the syndicate was illegal in Australia because it was “selling lottery tickets without a license” but many posters suggested it was also illegal due to being a pyramid scheme (not selling a real product as opposed to multi-level marketing which do sell real products) and offered bias personal anecdotes as to their experiences with the company. Those anecdotes were made by anonymous users on the forums.

As the hard drive owner had only been using the syndicate for a short time, had not signed anyone up and the data dated to 2003, and because it's unclear if it would have been illegal in New Zealand the decision was made not to contact the New Zealand Police. The drive was therefore classified as individually identifying, but not company identifying, as there were no company internal documents regarding the golf tracking system. The issues surrounding the syndicate meant it made more sense to err on the side of caution and not classify it as company identifying.

IDE_87 was another reformatted drive that appeared to have high school data on it. These data ranged from 2009 and 2010, and included quotes and bills to specific people and companies, and other high school specific data. The high school was not the same as any of the previously found high schools. Other individually identifying information was found as well.

IDE_88 was a clearly viewable drive that hosted a file server. It contained operating system and some data files but those files were non-identifying. Data carving yielded more files but again those were non-identifying.

IDE_89 was a clearly viewable drive. It had over 400 000 files on the hard drive making it one of the most populated hard drives. The hard drive predominantly had website development software on it and what appeared to be holiday images. Despite intensive searching nothing identifying was found, so data carving was used. This yielded over 5000 bitmaps and many other files. Examining those files also did not find any identifying information.

IDE_90 was a reformatted hard drive. It contained over 9000 graphics files when data carved. Those files were mainly high quality travel images and other artistic photographs. Nothing illicit or identifying was found. There did not appear to be any connection between IDE_89 and IDE_90.

IDE_91 was a reformatted drive with large amounts of data (over 400 000 files). The hard drive was clearly a high school drive and not related to any previously found high school drive. The data ranged from "Disk View" which had a list of user names and references to the high school's name to images that had the high school in the background, or religious themes which fitted the high school's denomination. Additionally one month's worth of bank records were found and whilst it is possible this was just a classroom assignment the numbers

appeared to be valid, real and logical for a school so were treated as real data. Other data identified individuals clearly and the data ranged from 2000 to 2006.

IDE_94 was a formatted drive. Data carving found a number of web based emails and documents in turn gave other identifying information about the users and owner of the hard drive. No company identifying information was found.

The fourth group comprised 6 IDE drives. All the drives were readable and all the drives reported back they had exactly 1 character that made up 100% of the space on each drive respectively. Therefore IDE_95, 96,97,98,99,100 were considered to be wiped. The auction description stated the drives were “Ex-lease” but did not include any indications if the drives had been formatted or wiped. In this instance that particular vendor most likely had wiped the drives.

Appendix D Ethical approval for this research

10/239

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

Academic Services
Manager, Academic Committees, Mr Gary Witte

17 December 2010

Dr H Wolfe
Department of Information Science
Division of Commerce
School of Business

Dear Dr Wolfe,

I am writing to let you know that, at its recent meeting, the Ethics Committee considered your proposal entitled "**Analysing the contents of Second Hand Hard Drives**".

As a result of that consideration, the current status of your proposal is:- **Approved**

For your future reference, the Ethics Committee's reference code for this project is:- **10/239**.

The comments and views expressed by the Ethics Committee concerning your proposal are as follows:-

The Committee appreciates the Category A application being made. The additional information provided has allayed initial concerns around personal information that resulted from the receipt of the Category B Reporting Sheet.

Approval is for up to three years. If this project has not been completed within three years from the date of this letter, re-approval must be requested. If the nature, consent, location, procedures or personnel of your approved application change, please advise me in writing.

Yours sincerely,

Gary Witte

Mr Gary Witte
Manager, Academic Committees
Tel: 479 8256
Email: gary.witte@otago.ac.nz

c.c. Assoc. Prof. M Winikoff Head Department of Information Science

Appendix E Valli's initial methodology

"The disk images were subjected to examination and experimentation with the methods listed below in increasing level of complexity until such time as documents and other digital artefacts on the drive were readily recoverable.

Level 1 - Mounting the drive and accessing the contents

This was accomplished by simply mounting the image as a volume on the investigative computer

Level 2 - Searching for Date/Day strings with a hexeditor/Undoing format of drive

1. Use of a simple hexeditor to search the image for common date strings such as Mon, Tues,16/3/2004, 2004.

2. The use of freely available unformat tools suited to the particular hard drive image format.

Level 3 - Interrogation of the image using "foremost" (quick mode)

Level 4 - Interrogation of the image using "foremost" (standard mode)

Level 5 - Analysis using Autopsy in attempt to recover documents"

(Valli, 2004, pg. 125)

Appendix F NIST definitions

NIST SP 800-88 provides four types of sanitisation activity: “Disposal”, “Clearing”, “Purging” and “Destroying”. As it is an American document, the American spelling (sanitization) has been left in, rather than adding [sic] after each occurrence.

Disposal is defined as:

> “Disposal is the act of discarding media with no other sanitization considerations. This is most often done by paper recycling containing non-confidential information but may also include other media.” (NIST, 2006, pg. 15)

Clearing is the process of:

> “Clearing information is a level of media sanitization that would protect the confidentiality of information against a robust keyboard attack. Simple deletion of items would not suffice for clearing. Clearing must not allow information to be retrieved by data, disk, or file recovery utilities.” Pg. 15-16 (NIST, 2006, pp. 15-16)

Purging is described as:

> “Purging information is a media sanitization process that protects the confidentiality of information against a laboratory attack. ... A laboratory attack would involve a threat with the resources and knowledge to use nonstandard systems to conduct data recovery attempts on media outside their normal operating environment. This type of attack involves using signal processing equipment and specially trained personnel.”
>
> (NIST, 2006, pg. 16)

The description for purging also includes a description of degaussing as it is considered a suitable form of purging. Degaussing is described as:

> “Degaussing is exposing the magnetic media to a strong magnetic field in order to disrupt the recorded magnetic domains. A degausser is a device that generates a magnetic field used to sanitize magnetic media.” (NIST, 2006, pg. 16)

Destruction is:

> “Destruction of media is the ultimate form of sanitization [sic]. After media are destroyed, they cannot be reused as originally intended. Physical destruction can be

accomplished using a variety of methods, including disintegration, incineration, pulverizing, shredding, and melting.

If destruction is decided upon due to the high security categorization of the information or due to environmental factors, any residual medium should be able to withstand a laboratory attack." (NIST, 2006, pg. 16)

Appendix G Vendor description of *Tableau*.

> "The Tableau is "a "forensic duplicator". As a forensic duplicator, the TD1 has many of the functions traditionally found in duplicators for the general IT market. In addition, the TD1 has features and capabilities that make it very good at handling the special needs of forensic practice.
> Like any good IT duplicator, the TD1 is *very* fast, sustaining data rates up to 6 GB/minute. The TD1 is also versatile, having native support for both SATA and IDE hard disks on both the input (source) and output (destination) interfaces.
> The TD1 also has features uniquely valuable in forensic applications. One of the most important of these features is the ability to calculate MD5 and SHA-1 hash values – sometimes called fingerprints – for the data being duplicated *without slowing the duplicator*. Other forensic features include detailed log generation (useful for case documentation), automatic blank-checking of source and destination drives, detection and handling of hidden/protected data areas on source and destination drives (HPA & DCO support), and so forth."

(Tableau LLC, 2009, pg. 4)

Appendix H ANZSIC LEVEL 1 CLASSIFICATIONS

These classifications were sourced from *The Australian and New Zealand Standard Industrial Classification (ANZSIC) 2006* (Stats New Zealand, 2006)

A	Agriculture, Forestry and Fishing
B	Mining
C	Manufacturing
D	Electricity, Gas, Water and Waste Services
E	Construction
F	Wholesale Trade
G	Retail Trade
H	Accommodation and Food Services
I	Transport, Postal and Warehousing
J	Information Media and Telecommunications
K	Financial and Insurance Services
L	Rental, Hiring and Real Estate Services
M	Professional, Scientific and Technical Services
N	Administrative and Support Services
O	Public Administration and Safety
P	Education and Training
Q	Health Care and Social Assistance
R	Arts and Recreation Services
S	Other Services
T	Not Elsewhere Included

Appendix I Consortium results 2006-2009

* Values were not reported for this field

Table 28 2006 Consortium Results (Jones, et al., 2006)

Country	Total	Unreadable	Wiped	Remaining	Company Identifying	Individual Identifying	Formatted	Illicit
UK	200	82 41%	37 31%	81	25 31%	24 30%	*	12 15%
North America	24	12 50%	1 8%	11	5 45%	6 55%	*	3 27%
Germany	40	29 72%	5 45%	6	2 33%	2 33%	*	*
Australia	53	3 6%	18 36%	32	14 44%	9 28%	*	2 6%

Table 29 2007 Consortium Results (Jones, Valli, Dardick, et al., 2009)

Country	Total	Unreadable	Wiped	Remaining	Company Identifying	Individual Identifying	Formatted	Illicit
UK	133	59 44%	28 38%	46	19 41%	30 65%	*	8 17%
North America	46	15 33%	6 19%	25	8 32%	15 60%	*	8 26%
Germany	42	30 71%	5 42%	7	2 29%	4 57%	*	*
Australia	79	8 10%	23 32%	48	22 46%	7 15%	*	6 8%

Table 30 2008 Consortium Results (Jones, Dardick, et al., 2009)

Country	Total	Unreadable	Wiped	Remaining	Company Identifying	Individual Identifying	Formatted	Illicit
UK	160	73 46%	27 31%	60	29 48%	32 53%	47 78%	4 5%
North America	63	7 11%	3 5%	53	11 21%	9 17%	14 26%	8 13%
Germany	28	11 39%	9 53%	8	2 25%	5 62%	6 75%	0 0%
France	44	22 50%	9 41%	13	5 38%	5 38%	9 69%	4 18%
Australia	43	6 14%	13 34%	25	15 60%	4 16%	7 28%	3 12%

Table 31 2009 Consortium Results (Jones, et al., 2010)

Country	Total	Unreadable	Wiped	Remaining	Company Identifying	Individual Identifying	Formatted	Illicit
UK	174	60 34%	32 28%	82	28 34%	36 45%	18 22%	2 2%
North America	74	13 18%	23 38%	38	14 37%	18 47%	10 26%	6 10%
Germany	39	8 21%	16 41%	15	4 26%	3 20%	3 20%	*
France	17	8 47%	4 24%	5	1 20%	1 20%	3 60%	*
Australia	42	5 12%	14 33%	23	18 78%	6 26%	*	*

Appendix J Spike Effect Systematic Sampling processing times

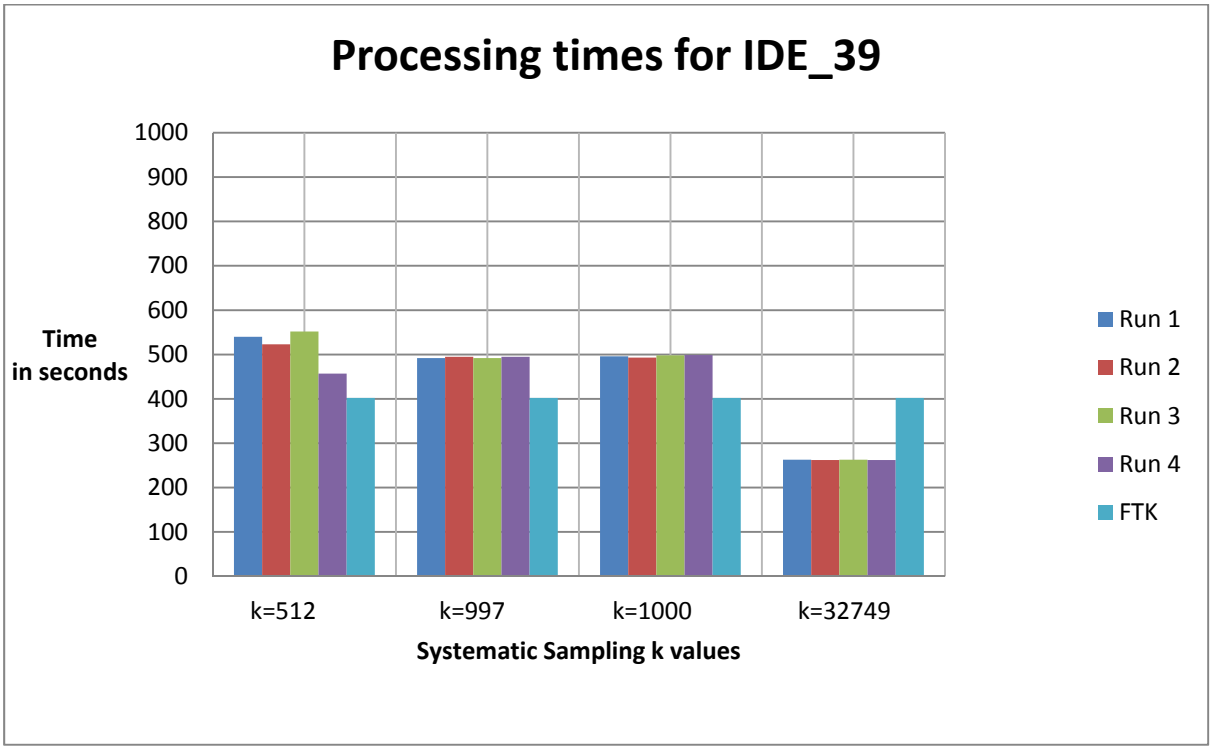


Figure 13 Comparison of processing times for IDE_39

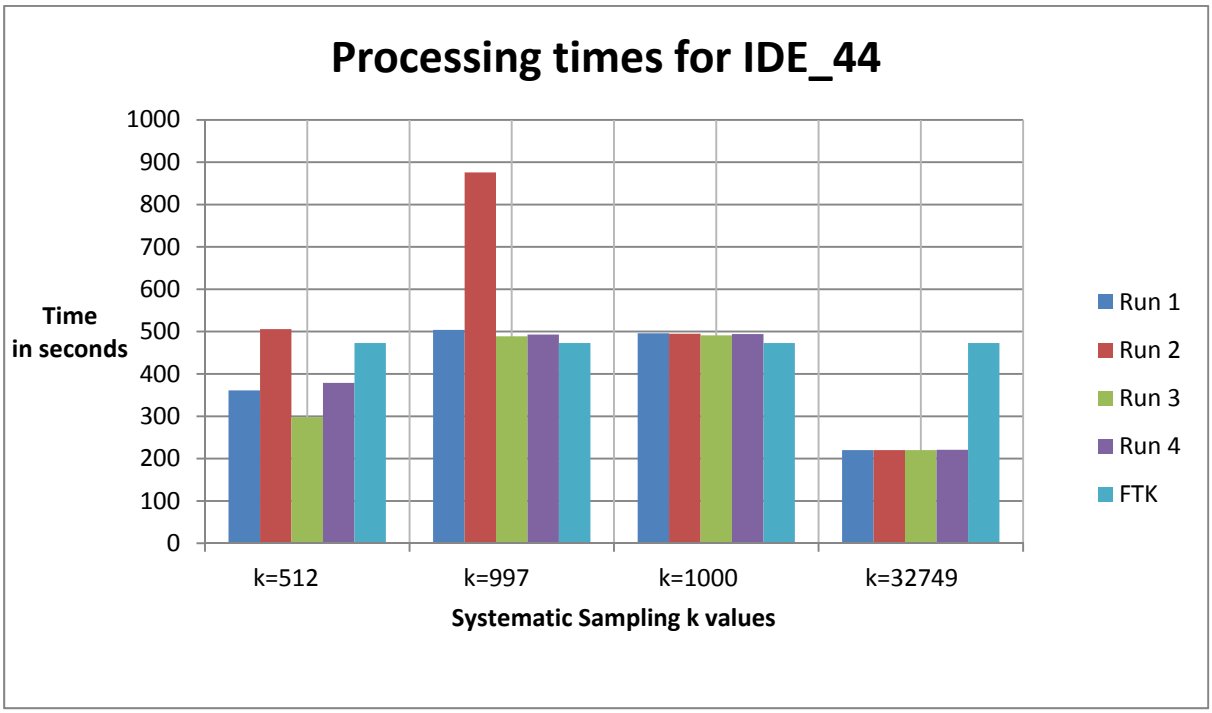


Figure 14 Comparison of processing times for IDE_44

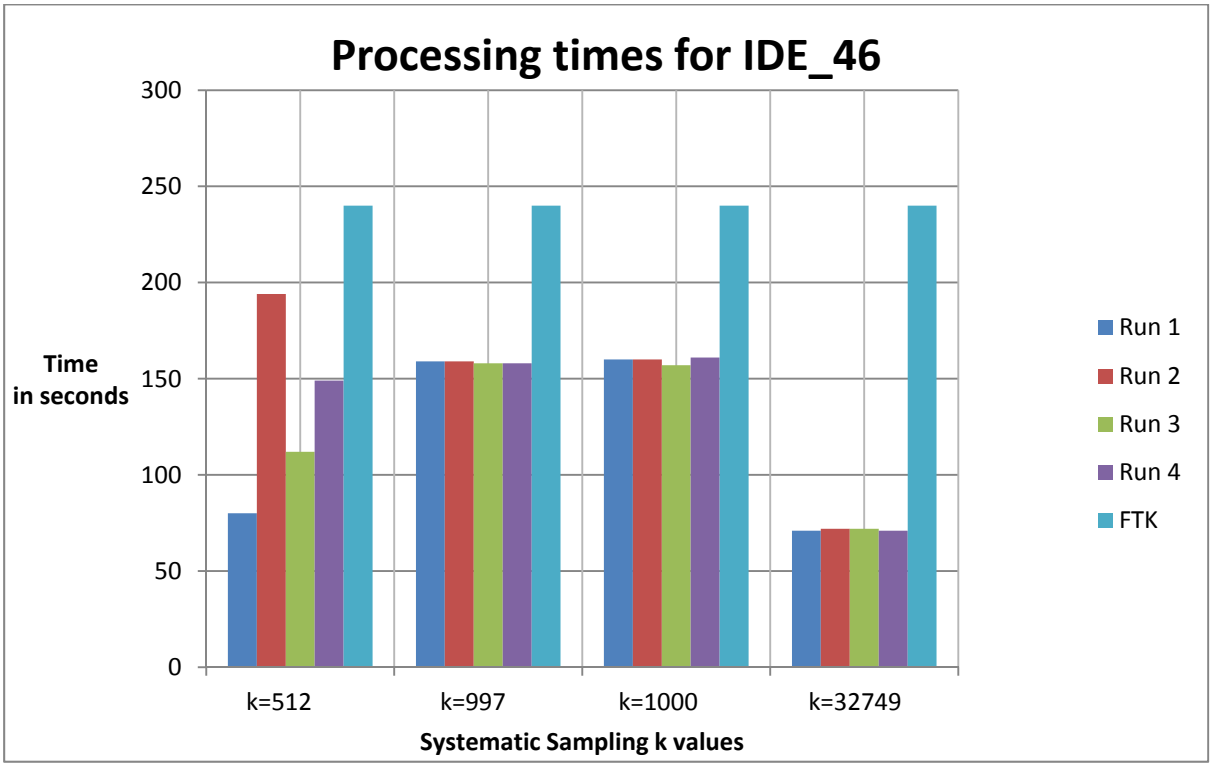


Figure 15 Comparison of processing times for IDE_46

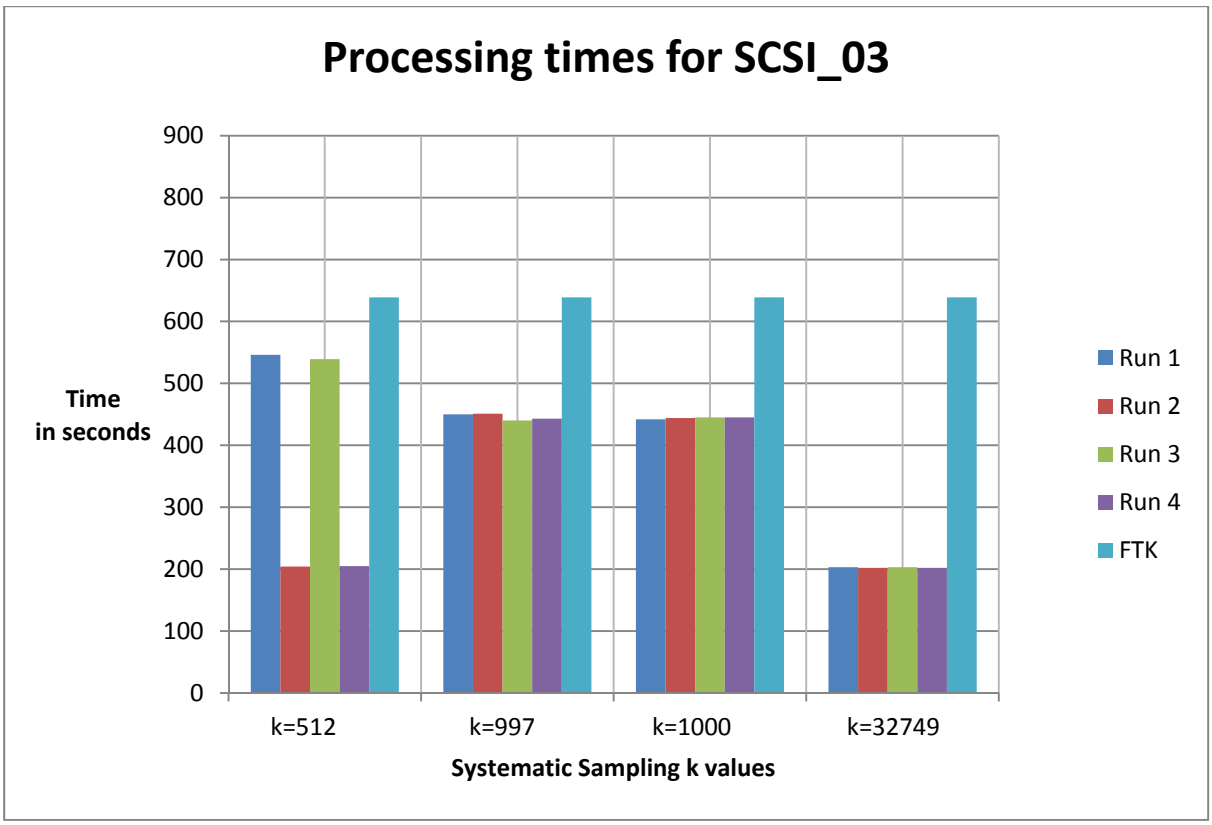


Figure 16 Comparison of processing times for SCSI_03

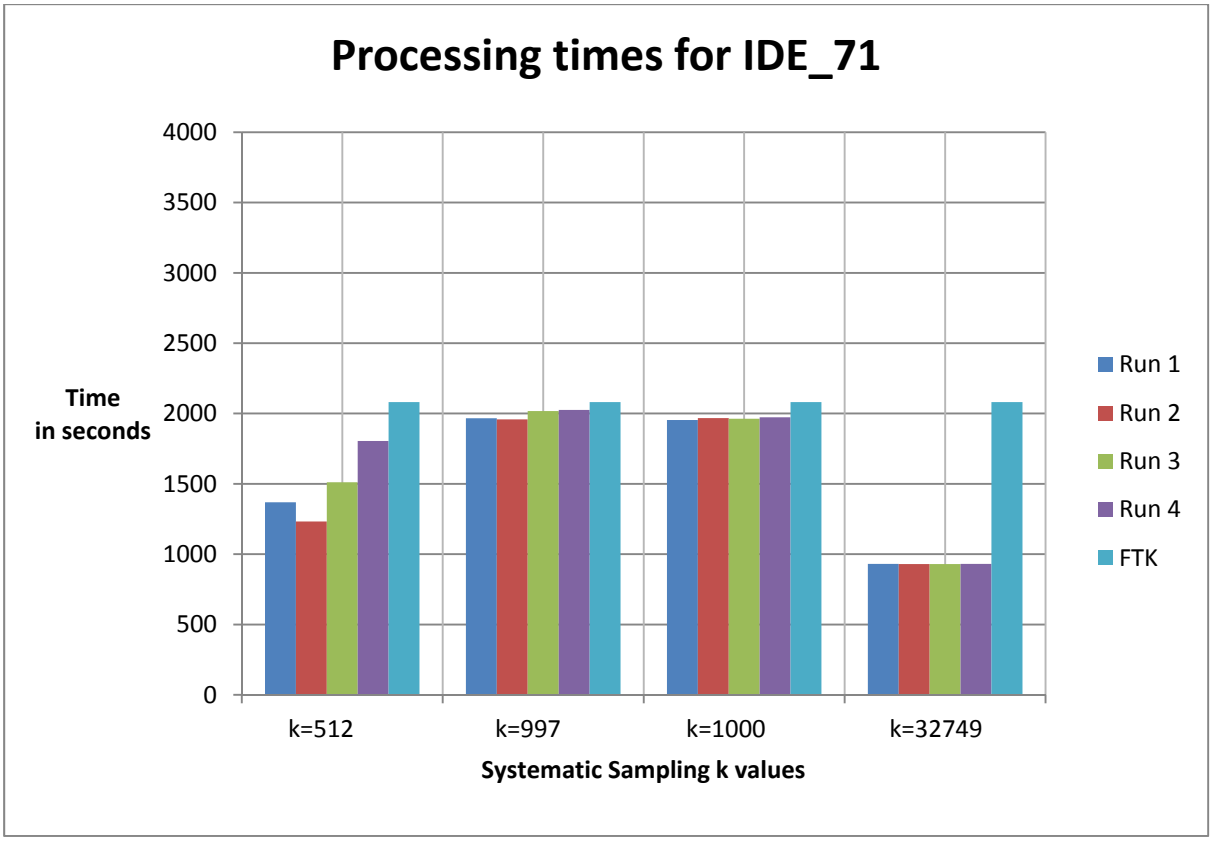


Figure 17 Comparison of processing times for IDE_71

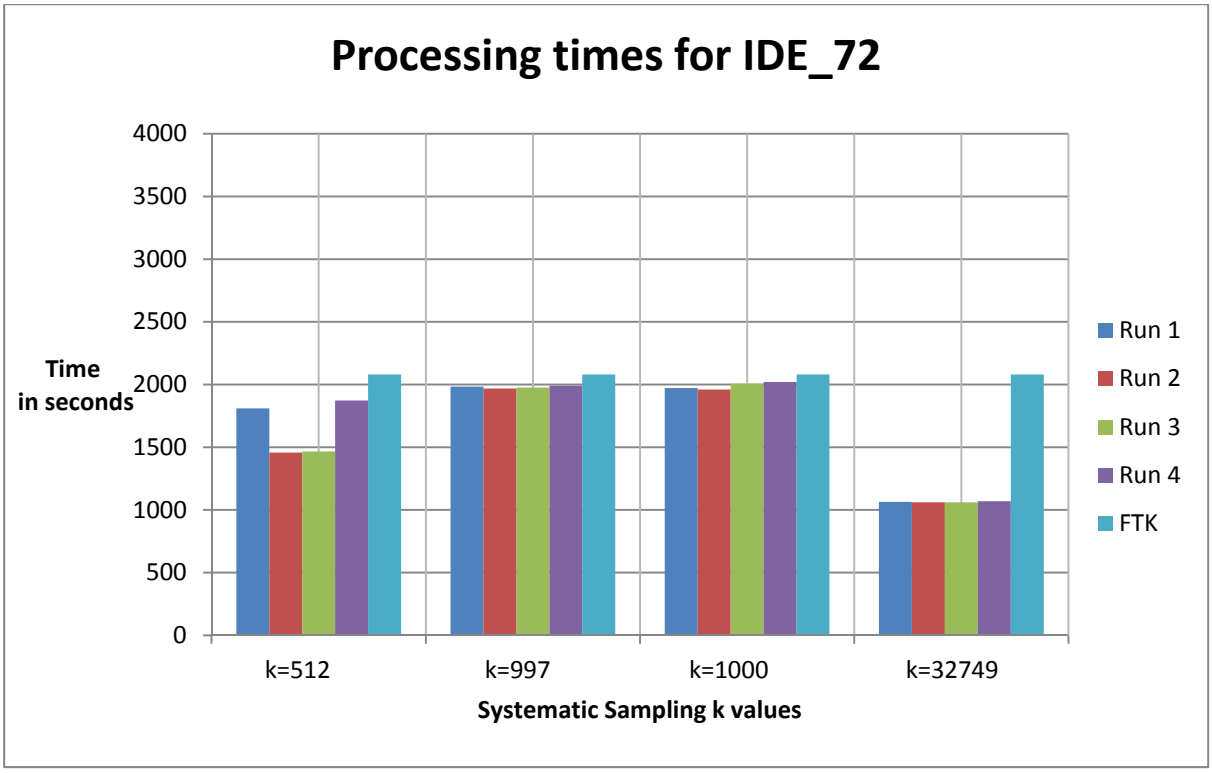


Figure 18 Comparison of processing times for IDE_72

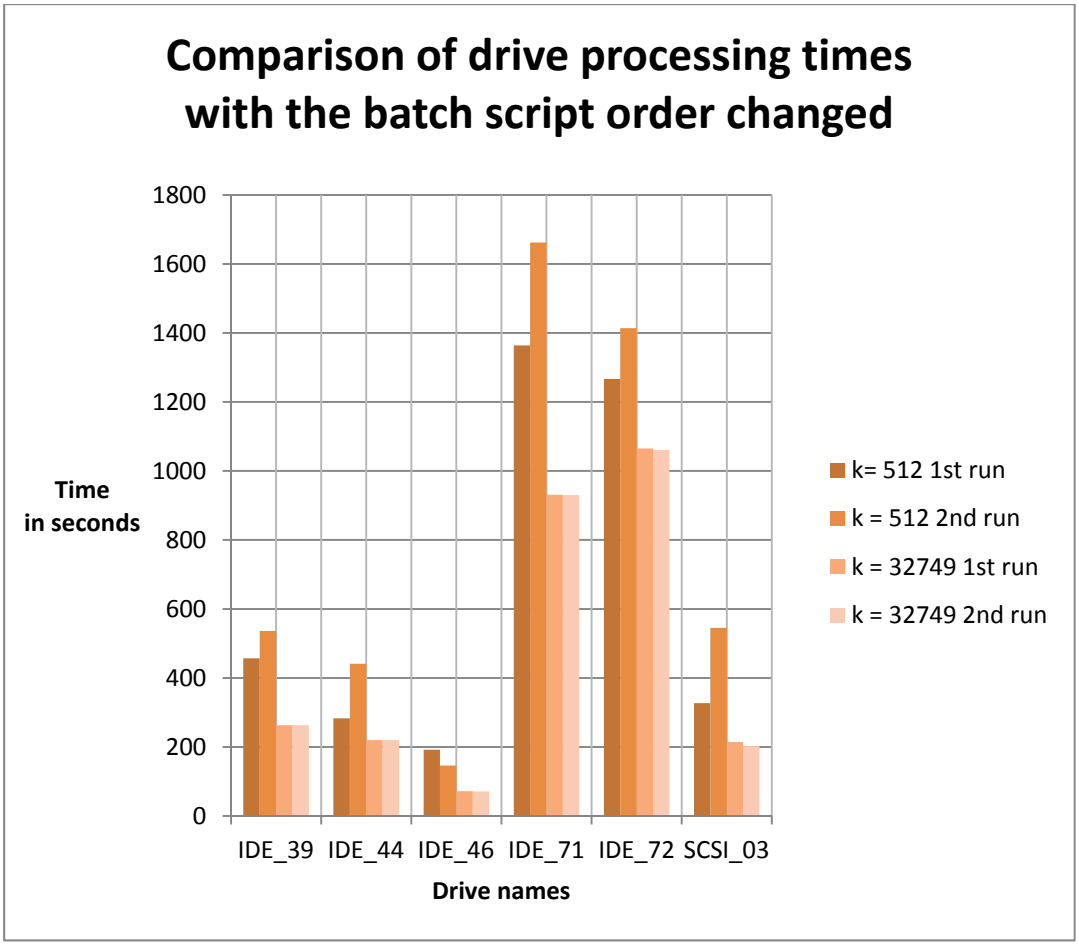


Figure 19 Comparison of processing times when the batch script has been changed

Appendix K Ranked comparisons of data remanence

Unreadable Drives : The lower the percentage the better as it means more drives remained to be studied.

Table 32 Rankings based on Unreadable Hard drives

Unreadable Drives	**Country**
12% (5)	Australia
18% (13)	North America
21% (21)	**New Zealand**
21% (8)	Germany
34% (60)	United Kingdom
47% (8)	France

Wiped Drives: The higher the percentage the better as it means those drives have been properly disposed of.

Table 33 Rankings based on Wiped Hard Drives

Wiped Drives	**Country**
43% (34)	**New Zealand**
41% (16)	Germany
38% (23)	North America
33% (14)	Australia
28% (32)	United Kingdom
24% (4)	France

Company Identifying Information: The lower the percentage the better as it means less information was found.

Table 34 Rankings based on Company Identifying Information

Company Identifying Information	Country
20% (1)	France
26% (4)	Germany
34% (28)	United Kingdom
37% (14)	North America
42% (19)	**New Zealand**
78% (18)	Australia

Individual Identifying Information: The lower the percentage the better as it means less information was found.

Table 35 Rankings based on Individual Identifying Information

Individual Identifying Information	Country
20% (3)	Germany
20% (1)	France
26% (6)	Australia
42% (19)	**New Zealand**
45% (36)	United Kingdom
46% (18)	North America